Multi-stage Classification of Diabetic Retinopathy and Age-related Macular Degeneration in Fundus Images Using CNNs and Vision Transformers

Dissertation submitted to
Visvesvaraya National Institute of Technology, Nagpur
in partial fulfillment of the requirements for the award of
the degree

Master of Technology
in
Computer Science & Engineering

by

Ayush Garg (MT23MCS013)

under the guidance of

Dr. Meera M. Dhabu

Department of Computer Science & Engineering
Visvesvaraya National Institute of Technology
Nagpur 440 010 (India)

JUNE 2025

Multi-stage Classification of Diabetic Retinopathy and Age-related Macular Degeneration in Fundus Images Using CNNs and Vision Transformers

Dissertation submitted to
Visvesvaraya National Institute of Technology, Nagpur
in partial fulfillment of the requirements for the award of
the degree

Master of Technology
in
Computer Science & Engineering

by

Ayush Garg (MT23MCS013)

under the guidance of
Dr. Meera M. Dhabu

Department of Computer Science & Engineering
Visvesvaraya National Institute of Technology
Nagpur 440 010 (India)

JUNE 2025

©Visvesvaraya National Institute of Technology (VNIT) 2025

Department of Computer Science & Engineering
Visvesvaraya National Institute of Technology, Nagpur

Declaration

I, **Ayush Garg**, hereby declare that this dissertation titled **"Multi-stage Classification of Diabetic Retinopathy and Age-related Macular Degeneration in Fundus Images Using CNNs and Vision Transformers"** is carried out by me in the Department of Computer Science & Engineering of Visvesvaraya National Institute of Technology, Nagpur. The work is original and has not been submitted earlier whole or in part for the award of any degree at this or any other Institution.

Ayush Garg
Name and signature
Date: 16-06-2024

Certificate

This to certify that the dissertation titled **"Multi-stage Classification of Diabetic Retinopathy and Age-related Macular Degeneration in Fundus Images Using CNNs and Vision Transformers"** is submitted by **Ayush Garg** in partial fulfillment of the requirements for the award of the degree of **Master of Technology in Computer Science & Engineering**, VNIT Nagpur. The work is comprehensive, complete, and fit for final evaluation.

Dr. Meera M. Dhabu
Department of Computer Science &
Engineering, VNIT Nagpur

Dr. M. P. Kurhekar
Head, Department of Computer Science
& Engineering, VNIT Nagpur
Date:

विभाग प्रमुख
Head
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Dept. of Computer Science & Engg.
वि.रा.प्रौ.सं. नागपुर-१० (भारत)
V.N.I.T., Nagpur-440010. (India)

Department of Computer Science & Engineering
Visvesvaraya National Institute of Technology, Nagpur

Non-Plagiarism Certificate

Certified that M. Tech Dissertation titled **"Multi-stage Classification of Diabetic Retinopathy and Age-related Macular Degeneration in Fundus Images Using CNNs and Vision Transformers"** submitted by **Ayush Garg**, Enrollment No. **MT23MCS013**, ID No. **30538** has been checked for plagiarism using TURNITIN software and the overlap is found within prescribed limits. A summary of the report is given below.

ORIGINALITY REPORT

15%	12%	10%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

Ayush Garg
Enrollment No: MT23MCS013

Prof. Meera M. Dhabu
Professor
Department of Computer Science & Engineering
VNIT, Nagpur

ACKNOWLEDGEMENT

I would like to express my deepest gratitude and sincere appreciation to all those who have supported and guided me throughout this M.Tech journey.

First and foremost, I am immensely grateful to my supervisor, **Dr. Meera M. Dhabu** Madam, for her invaluable guidance, expertise, and encouragement. Her constant support, patience, and insightful feedback have been instrumental in shaping this thesis. I am truly fortunate to have the opportunity to work under her mentorship.

I would like to extend my heartfelt thanks to **Dr. M. P. Kurhekar**, Head of Department, for his invaluable support and guidance throughout this journey. His encouragement has been vital in the successful completion of this project.

I would like to extend my heartfelt thanks to the faculty members of the Department of Computer Science & Engineering at Visvesvaraya National Institute of Technology Nagpur for their knowledge-sharing and for providing a conducive academic environment. I am indebted to them for their teachings and guidance which have broadened my perspective and enhanced my understanding of the subject. My sincere appreciation goes to them for their assistance, collaboration, and providing necessary resources for conducting experiments and carrying out this research work. Their contributions have been significant in the successful completion of this thesis.

I would also like to express my gratitude to my family and friends for their unwavering support, understanding, and encouragement throughout this journey. Their love, motivation, and belief in my abilities have been the driving force behind my accomplishments.

Thank you all.

Ayush garg

Abstract

In India, we're witnessing a troubling rise in retinal diseases—most notably Diabetic Retinopathy (DR) and Age-related Macular Degeneration (AMD) —both of which are major contributors to vision loss. AMD manifests in two distinct forms— Dry and Wet—each demanding a different therapeutic strategy. Early detection and accurate classification of these conditions are therefore crucial for ensuring timely, effective treatment. Yet most automated systems today either distinguish only between diseased and healthy retinas or focus on a single disorder, leaving them ill-equipped for multi-disease screening.

To address this shortfall, we propose a hierarchical, two-stage pipeline that leverages Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs). In the first stage, the proposed model coarsely sorts images into three groups—DR, AMD (Dry + Wet), and healthy retinas—allowing it to learn broad distinctions on a balanced dataset. The second stage then zooms in on the AMD cases, splitting them into Dry versus Wet AMD. By decomposing the task into two focused steps, the proposed model minimizes confusion between very different disease types and yields more reliable predictions.

We evaluated our pipeline on a carefully curated set of fundus photographs and found that this multi-stage approach consistently outperforms traditional one-shot classifiers in accuracy, precision, recall, and F1-score. Our results demonstrate a scalable, real-world solution for AI-powered retinal screening—one that can seamlessly integrate into busy clinics and resource-limited settings while improving diagnostic accuracy across multiple retinal diseases.

TABLE OF CONTENTS

Title **Page No.**

ABSTRACT **iv**

LIST OF TABLES **vii**

LIST OF FIGURES **viii**

1 Introduction **1**

1.1 Clinical Background 3

1.2 Retinal Imaging Techniques 5

1.3 Motivation 8

1.4 Motivation Behind the Two-Stage Design 9

1.5 Challenges in Retinal Image Processing 10

2 Literature Review **11**

2.1 AMD Classification 11

2.1.1 Approch using Fundus Image for classification 11

2.1.2 Model Using Both Fundus and OCT Images 14

2.2 Diabetic Retinopathy Classification 14

2.2.1 Traditional Machine Learning Methods 14

2.2.2 Deep Learning Approaches (CNNs) 15

2.2.3 Transformer and Advanced Architectures 16

3 Proposed Methodology **18**

3.1 Dataset Preparation 18

3.2 Image Preprocessing and Augmentation 21

3.3 Cascade Framework 21

3.4 Implemented Model Architectures 22

3.4.1 ResNet-50 Wide 22

3.4.2 ResNet-50 23

3.4.3 Inception-V4 24
3.4.4 EfficientNetV2 28
3.4.5 Visformer 28
3.4.6 DeiT-Tiny Patch16-224 30
3.4.7 ViT-Tiny Patch16-224 30
3.4.8 BEiT-Large Patch16-224 31
3.5 Training Parameters and Implementation Details 31

4 Evaluation Matrix **34**

5 Results and Discussion **36**
5.1 Stage 1: Normal vs DR vs AMD Classification 36
5.2 Stage 2: Dry AMD vs Wet AMD Classification 36
5.3 Overall Discussion And Conclusion 41
5.4 Comparative Bar-Graph Analysis 44
5.5 Conclusion 50

6 Future Work **51**

List of Tables

5.1 Performance of CNN-based models for Stage 1: Normal vs DR vs AMD classification 37
5.2 Performance of Transformer-based models for Stage 1: Normal vs DR vs AMD classification 38
5.3 Performance of CNN-based models for Stage 2: DRY vs WET classification 39
5.4 Performance of Transformer-based models for Stage 2: DRY vs WET classification 40
5.5 Stage 1 (Normal vs DR vs AMD) — CNN-based Models Summary . . . 42
5.6 Stage 1 (Normal vs DR vs AMD) — Transformer-based Models Summary 42
5.7 Stage 2 (DRY vs WET) — CNN-based Models Summary 42
5.8 Stage 2 (DRY vs WET) — Transformer-based Models Summary 43

List of Figures

1.1 Yellow lipid-rich drusen deposits in Dry AMD 3
1.2 Choroidal neovascularization seen in Wet AMD 4
1.3 Cross-sectional schematic of the eye, highlighting key structures: retina, choroid, macula, optic disc, and vitreous chamber. 6
1.4 Color fundus photograph showing the optic disc (OD), macula (M), and retinal vasculature. 7

3.1 **Sample Fundus Images for Each Class.** Representative examples of (a) Diabetic Retinopathy, (b) Normal Retina, (c) Dry AMD, and (d) Wet AMD used in our dataset. 20
3.2 Cascaded deep learning framework for fundus image classification. 22
3.3 The architecture of ResNet-50, consisting of convolutional layers grouped into four stages with residual connections. 25
3.4 Overview of the Inception-V4 architecture showing the sequence of modules and output shapes at each stage. 25
3.5 Detailed structure of Inception-A, Inception-B, and Inception-C modules used within Inception-V4. 26
3.6 Architecture of EfficientNet-B3 model showing the MBConv blocks, compound scaling, and squeeze-and-excitation mechanisms. 27
3.7 Architecture of the VIT model used for fundus image analysis. 29

5.1 Stage 1 Accuracy of CNN models 44
5.2 Performance on Stage 1 (Normal vs DR vs AMD) with CNN Models . . . 45
5.3 Stage 1 Accuracy of Transformer models 46
5.4 Performance on Stage 1 (Normal vs DR vs AMD) with Transformer Models 47
5.5 Performance on Stage 2 (Dry AMD vs Wet AMD) on CNN models 48
5.6 Performance on Stage 2 (Dry AMD vs Wet AMD) on Transformer model 49

Chapter 1

Introduction

The human eye is one of the most essential organs that helps us perceive and make sense of the world around us. It allows us to recognize shapes, colors, distances, and movement, all of which play a huge role in our daily activities, learning process, and overall quality of life. Good vision is key to maintaining independence, safety, and communication. Unfortunately, despite its importance, eye health often doesn't get the attention it deserves until a noticeable decline in vision occurs.

With the progress in medical science, many eye conditions can be detected and managed early. Yet, global data still shows that millions of people continue to suffer from visual impairments that could have been prevented or treated. According to the World Health Organization (WHO), over 2.2 billion people globally live with some form of vision impairment or blindness. Shockingly, at least 1 billion of those cases could have been avoided or still remain unaddressed. The causes range from easily correctable issues like refractive errors and cataracts to more complex retinal diseases such as Age-related Macular Degeneration (AMD) and Diabetic Retinopathy (DR). These two, in particular, are among the top causes of irreversible vision loss.

AMD mostly affects older adults and damages the macula, the central part of the retina responsible for sharp, detailed vision. It is mainly classified into two types: Dry AMD and Wet AMD. Dry AMD is marked by thinning of the macula and the presence of tiny protein deposits called drusen. Wet AMD, on the other hand, occurs when abnormal blood vessels grow beneath the retina, leaking fluid and causing further damage.

In developed nations, AMD is the leading cause of severe and permanent vision loss in people over the age of 60. WHO data shows it accounts for a significant share of vision impairment in high-income countries. Currently, more than 196 million people around the world are living with AMD, and this number is projected to reach 288 million by 2040 due to increasing life expectancy.

Diabetic Retinopathy (DR) is a diabetes-related complication that affects the blood

vessels in the retina. Consistently high blood sugar levels over time damage these vessels, making them swell, leak, or even close off completely. DR advances in stages, starting from mild nonproliferative changes to the more dangerous proliferative stage, which can result in retinal detachment and permanent vision loss.

According to WHO, DR accounts for around 4.8% of global blindness. The likelihood of developing DR increases with the duration of diabetes. Nearly all individuals with type 1 diabetes and over 60% of those with type 2 diabetes are expected to develop some form of retinopathy after living with the disease for 20 years.

In India, the number of people affected by retinal conditions is rising quickly, driven by a growing elderly population and the increasing prevalence of diabetes. As per the National Diabetes and Diabetic Retinopathy Survey (2015–2019) conducted by the Ministry of Health and Family Welfare, the prevalence of DR in India stands at 16.9% among known diabetic patients. With more than 77 million people currently living with diabetes in India, this translates to over 13 million individuals potentially at risk of DR.

Age-related Macular Degeneration is also becoming a public health concern in India, especially among the elderly. Recent population-based studies estimate that early and late stages of AMD affect approximately 1.38% and 0.32% of Indians aged 60 and above, respectively. With India projected to have 30 million senior citizen by 2050, the prevalence of AMD is expected to increase significantly

Detecting retinal diseases like AMD and DR early is crucial to preventing irreversible damage. Thanks to advancements in imaging technology, deep learning, and artificial intelligence, automated diagnosis from retinal fundus images is now becoming a reality. These tools can help eye specialists screen large groups of people more efficiently and accurately.

Timely intervention can slow down disease progression and help retain vision. For example, anti-VEGF treatments have shown positive results in managing Wet AMD, while early laser therapy or vitrectomy can prevent serious vision loss in advanced DR. However, the effectiveness of these treatments relies heavily on how early the disease is identified.

The eye is more than just a window to the world—it is also a mirror of systemic health. Many systemic diseases first show signs in the retina. The growing number of retinal disease cases globally and in India makes it more important than ever to raise awareness, encourage regular eye checkups, and promote timely treatment. By harnessing the power of modern diagnostic tools and deep learning models for fundus image analysis, we can take strong steps toward reducing visual impairment and protecting the gift of sight for millions.

1.1 Clinical Background

Age-related Macular Degeneration (AMD) and Diabetic Retinopathy (DR) are two of the most impactful retinal diseases today, significantly affecting vision health worldwide. These conditions can lead to partial or even complete blindness and often have a profound effect on the quality of life of those affected. Having a solid understanding of their clinical features is essential, especially when developing automated diagnostic models.

AMD mainly affects the macula—the central region of the retina that's responsible for sharp, color vision needed for everyday tasks like reading, driving, and recognizing faces. Clinically, AMD is divided into two main categories: **Dry AMD** and **Wet AMD**. Dry AMD is the more common type and is typically associated with the accumulation of drusen—yellow, lipid-rich deposits that form between the retinal pigment epithelium (RPE) and Bruch's membrane, as shown in Figure 1.1. These deposits gradually lead to thinning and degeneration of macular tissue, resulting in progressive central vision loss.

Wet AMD, while less common, is much more aggressive. It's characterized by the growth of abnormal blood vessels under the retina—a process known as choroidal neovascularization (CNV). These vessels tend to leak fluid or bleed, damaging the retinal cells and often leading to permanent vision loss if left untreated (see Figure 1.2).

Figure 1.1: Yellow lipid-rich drusen deposits in Dry AMD

For clinical diagnosis, ophthalmologists often rely on imaging tools like fundus photography, fluorescein angiography (FA), and optical coherence tomography (OCT). OCT

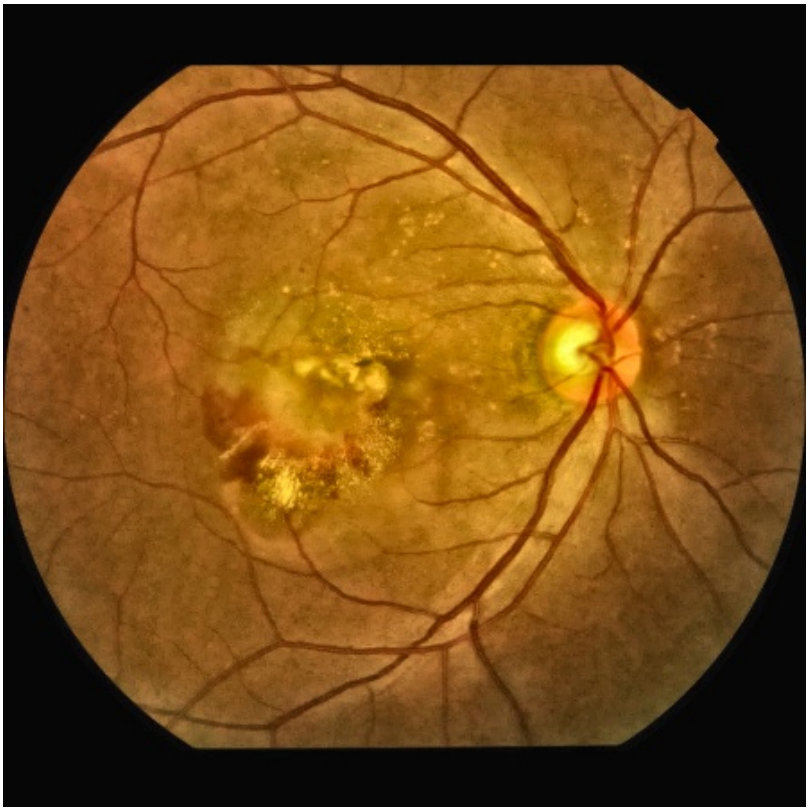


Figure 1.2: Choroidal neovascularization seen in Wet AMD

is especially useful as it provides detailed cross-sectional views of the retina, helping in the identification of drusen, fluid buildup, and CNV.

Diabetic Retinopathy (DR), on the other hand, is a microvascular complication of long-term diabetes. Prolonged high blood sugar damages the tiny vessels in the retina, leading to leakage, blockage, and eventually ischemia. DR progresses through well-defined stages—from mild non-proliferative DR (NPDR), which includes microaneurysms, to moderate and severe NPDR, marked by hemorrhages, cotton wool spots, and venous abnormalities. The most advanced stage is proliferative DR (PDR), characterized by the growth of new blood vessels (neovascularization), which can cause serious complications like vitreous hemorrhage or retinal detachment.

Key clinical signs of DR include:

- **Microaneurysms:** Tiny red dots and the earliest observable feature.
- **Dot and blot hemorrhages:** Result from ruptured microaneurysms.
- **Hard exudates:** Fatty deposits due to leaky blood vessels.
- **Cotton wool spots:** Small white patches caused by localized retinal nerve fiber infarctions.

- **Neovascularization:** New blood vessel growth, usually seen in PDR stages.

To diagnose DR accurately, doctors typically perform a dilated fundus exam, capture color fundus images, and use OCT scans. Fundus fluorescein angiography (FFA) is also important in evaluating vascular leakage and areas of non-perfusion.

Clinical Trials and Experiments: Several major studies have helped shape our understanding of AMD and DR. The Age-Related Eye Disease Study (AREDS) and its follow-up, AREDS2, highlighted the role of nutritional supplements in slowing AMD progression. For DR, studies like the Diabetes Control and Complications Trial (DCCT) and the UK Prospective Diabetes Study (UKPDS) have clearly shown how tight control of blood sugar levels can delay the onset and slow the progression of diabetic eye disease.

Current Treatment Approaches:

- **For AMD:** Anti-VEGF injections—such as ranibizumab and aflibercept—are the go-to treatment for wet AMD. For dry AMD, there is no definitive treatment yet, but the AREDS vitamin formula can help slow down progression in many patients.

- **For DR:** Treatments vary based on the stage but commonly include laser photocoagulation, anti-VEGF therapies, corticosteroids, and vitrectomy for advanced cases.

The Role of Imaging in AI Models: The availability of high-quality, annotated retinal images from hospitals and clinical trials has made it possible to train deep learning models for disease detection and grading. These AI systems rely heavily on accurate labels and expert annotations to match or even exceed the diagnostic performance of trained ophthalmologists.

Integrating artificial intelligence into eye care systems can make a huge difference, especially in under-resourced areas where access to specialists is limited. AI can help scale screening efforts, reduce the burden on clinicians, and make early diagnosis more accessible to the general population.

1.2 Retinal Imaging Techniques

Accurate imaging of the retina is indispensable for diagnosing and tracking diseases like Diabetic Retinopathy (DR) and Age-related Macular Degeneration (AMD). By peering into the eye's interior, clinicians can catch early signs of trouble, monitor changes over time, and tailor treatments to each patient's needs. Although several methods exist, two stand out in routine practice: **fundus photography** and **Optical Coherence Tomography (OCT)**. Advanced techniques such as fluorescein angiography and the newer OCT-Angiography build on these foundations.

Fundus Photography

Fundus photography offers a simple, non-invasive way to capture vivid, two-dimensional snapshots of the retina, optic disc, macula, and blood vessels. Using a low-power microscope fitted with a camera, this method produces images similar to the view shown in Figure 1.4—a direct projection of the retinal surface depicted in the schematic cross-section of Figure 1.3.

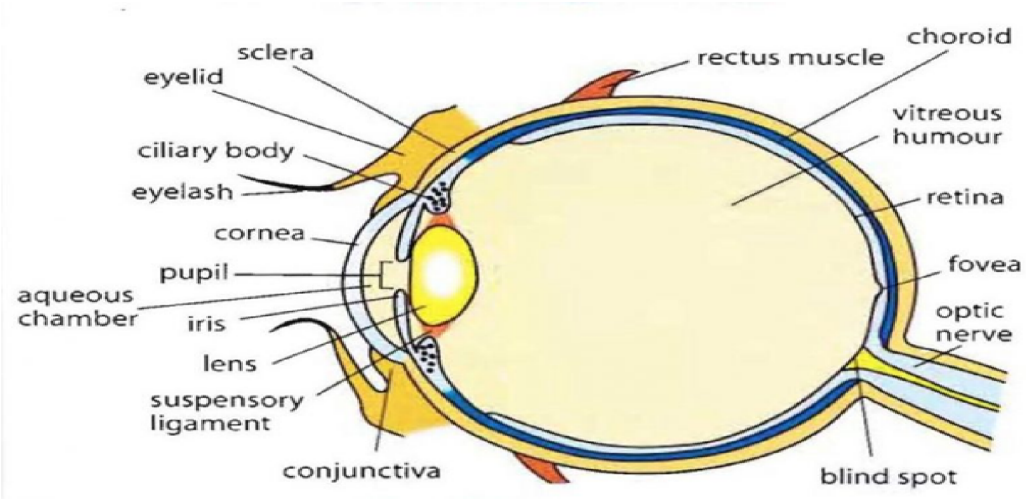


Figure 1.3: Cross-sectional schematic of the eye, highlighting key structures: retina, choroid, macula, optic disc, and vitreous chamber.

This technique shines in screening programs, especially where resources are tight:

- It's swift and well tolerated by patients.
- It reveals a broad view of retinal health, flagging hemorrhages, microaneurysms, drusen, and more.
- Its cost-effectiveness makes it ideal for widespread diabetic retinopathy screening in low- and middle-income regions.
- On the downside, it doesn't show depth, so subtle, layer-specific changes remain hidden, and image clarity can suffer when cataracts or patient movement come into play.

Optical Coherence Tomography (OCT)

OCT brings depth to retinal imaging. Using low-coherence light waves, it slices through the layers of the retina, delivering high-resolution, cross-sectional views that are invaluable for spotting drusen, fluid pockets, and other fine-scale details.

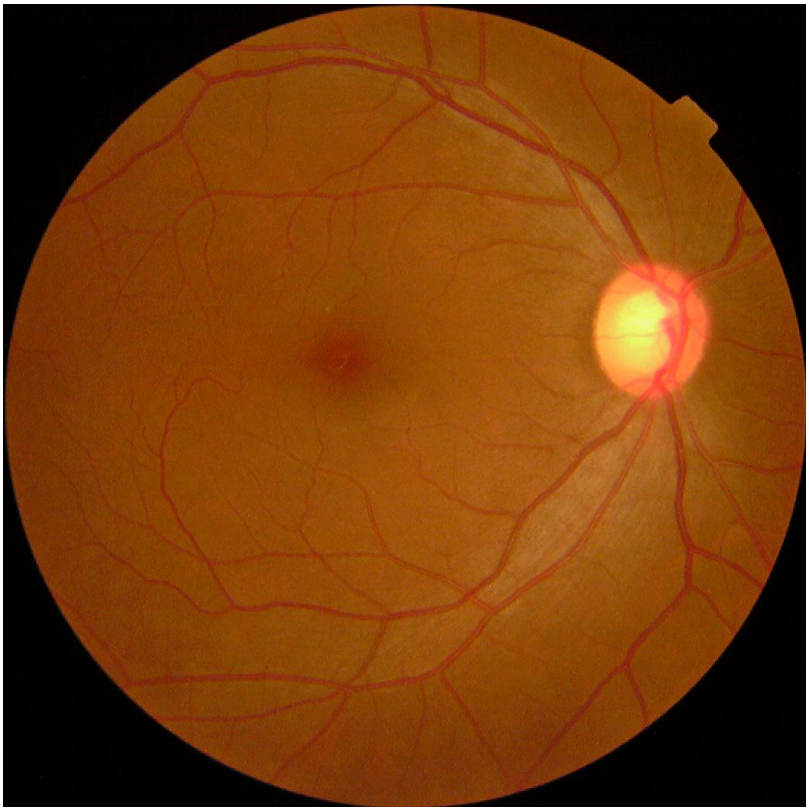


Figure 1.4: Color fundus photograph showing the optic disc (OD), macula (M), and retinal vasculature.

In AMD care, OCT helps:

- Pinpoint drusen deposits and areas of atrophy in the dry form.
- Track fluid accumulation and neovascular membranes in wet AMD.
- Gauge treatment response after anti-VEGF injections.

For DR, OCT can:

- Detect macular edema and subtle retinal thickening.
- Reveal vitreomacular traction that might worsen vision.

Of course, OCT systems come with a higher price tag, and effective use demands both patient cooperation and trained operators. Availability in remote clinics can be a limiting factor.

Fluorescein Angiography (FA)

When detailed vascular mapping is needed, fluorescein angiography steps in. After injecting fluorescein dye into a vein, a series of rapid-fire photographs captures the dye's journey

through retinal vessels, unveiling areas of leakage, capillary drop-out, and abnormal vessel growth. This approach is especially useful for confirming choroidal neovascularization in AMD or neovascularization in proliferative DR. However, its invasive nature comes with possible side effects—from mild nausea to, in rare cases, allergic reactions.

OCT Angiography (OCT-A)

The latest advance, OCT-A, merges the best of OCT and angiography without dye injections. By detecting motion within blood vessels, OCT-A sketches out detailed maps of retinal and choroidal circulation layer by layer. It excels at identifying early neovascular membranes and areas of nonperfusion, although its smaller field of view and lack of leakage information mean it complements rather than replaces FA.

Integrating AI into Retinal Imaging

The explosion of deep learning has ushered in new possibilities: automated analysis of fundus and OCT scans. Well-trained AI models can flag hallmark signs—hard exudates, microaneurysms, hemorrhages, drusen, fluid accumulations—and triage cases by urgency. In practice, these tools can relieve clinicians' workload, support large-scale screening in underserved areas, and help ensure that high-risk patients see specialists sooner.

Summary

No single imaging modality does it all. Fundus photography remains the first step in mass screening, prized for its simplicity and reach. OCT and OCT-A deepen our insights with structural and vascular detail. By weaving together these techniques—augmented by AI—we can catch retinal diseases earlier, tailor treatments more precisely, and ultimately preserve vision for more people across diverse healthcare settings.

1.3 Motivation

Interpreting retinal images is a nuanced skill honed through years of training. It's a meticulous, time-consuming process, prone to differences in interpretation between clinicians. In many parts of the world, there simply aren't enough specialists to keep pace with the rising tide of DR and AMD, leading to delayed diagnoses and preventable vision loss.

Imagine rural clinics without easy access to ophthalmologists—patients might wait months or travel hundreds of kilometers for a basic screening. By then, subtle signs may have progressed to irreversible damage. That's where AI-driven Computer-Aided Diagnosis (CAD) systems come into play. These platforms can rapidly sift through images,

highlight suspicious findings, and ensure that those who need care most get referred without delay.

Deploying portable fundus cameras or handheld OCT devices alongside AI algorithms can transform primary care settings. A nurse captures images, the AI flags any red flags, and ophthalmologists review only the critical cases. This model not only speeds up referrals but also boosts equity, bringing quality eye care to underserved communities. Because AI doesn't tire or lose focus, it provides consistent, unbiased assessments around the clock.

As healthcare systems digitize and cloud infrastructure expands, integrating AI into national eye care programs is not just feasible—it's essential. Our project aims to harness these technologies to bridge gaps in access, reduce preventable blindness, and enhance the efficiency of retinal disease management worldwide. "'

1.4 Motivation Behind the Two-Stage Design

In real-world retinal image datasets, some disease categories are heavily underrepresented—Dry AMD often has fewer than a couple of hundred examples, while Normal and DR images can number in the thousands. If we trained a single model to recognize all four classes (Normal, DR, Dry AMD, Wet AMD) at once, it would struggle to learn the subtle differences that make Dry AMD unique, simply because it sees so few examples of it.

To work around this, we split the problem into two steps, much like an ophthalmologist would. First, our model learns to separate Normal, DR, and AMD (where AMD includes both Dry and Wet cases).

Once an image is flagged as AMD, we pass it to a second model that specializes in telling Dry AMD apart from Wet AMD. This two-stage approach offers several advantages:

- **Balanced Learning:** Each model focuses on a smaller, more balanced set of classes, which improves overall accuracy.
- **Reduced Confusion:** We avoid asking one model to make fine-grained distinctions across very different conditions, cutting down on misclassifications.
- **Clinical Alignment:** It mirrors how practitioners triage cases—first deciding if a patient has AMD, then determining the subtype.
- **Practical Flexibility:** In low-resource settings where balanced data is hard to gather, this pipeline remains robust without needing equal numbers of examples for every category.

1.5 Challenges in Retinal Image Processing

Processing retinal fundus images presents a unique set of challenges that differ markedly from general image-processing tasks. First, the clinically relevant features in retinal images—microaneurysms, hemorrhages, exudates—are often subtle, low-contrast, and highly localized. Accurately detecting and segmenting these lesions demands exceptional precision and domain expertise, making off-the-shelf computer-vision methods difficult to apply directly.

A second challenge is the scarcity of large, well-annotated datasets. Labeling retinal images requires trained ophthalmologists, a time-consuming and expensive process. As a result, datasets are smaller and suffer from class imbalances—especially for rare disease stages—hindering the training of robust deep-learning models. By contrast, general-purpose repositories like ImageNet offer vast, easily annotated collections.

Retinal images also exhibit high variability in acquisition: different camera systems, lighting conditions, fields of view, and patient anatomy all introduce inconsistencies that complicate preprocessing and reduce model generalizability. Such variability is far less pronounced in standard object-recognition tasks.

Explainability poses yet another hurdle. Clinicians require models whose decisions can be interpreted and justified in medical terms. Unfortunately, many high-performance deep networks remain "black boxes," limiting their acceptance in clinical workflows without additional mechanisms for interpretability.

Finally, the frequent overlap of disease signs—and the possibility of comorbid conditions—demands fine-grained classification. A single fundus image might exhibit features of multiple retinal pathologies, increasing the risk of misdiagnosis in models that lack nuanced discrimination.

Together, these factors underscore the need for domain-tailored machine-learning pipelines, rigorous preprocessing, and close collaboration with ophthalmic experts to build accurate, trustworthy systems for retinal image analysis.

By chaining these two stages, we turn a daunting four-way classification into manageable, clinically meaningful steps—bringing us closer to reliable, field-ready retinal screening tools.

Chapter 2

Literature Review

This chapter provides a comprehensive overview of automated diagnostic methods for Age-related Macular Degeneration (AMD) and Diabetic Retinopathy (DR), two leading causes of vision loss globally. We explore diagnostic approaches using both fundus photography and Optical Coherence Tomography (OCT) modalities, trace the evolution from traditional machine learning with handcrafted features to modern deep learning techniques, and highlight gaps in unified multi-disease screening frameworks.

2.1 AMD Classification

Automated AMD detection methods are categorized by imaging modality: fundus photography, which is cost-effective and widely accessible, and OCT, which offers detailed cross-sectional retinal imaging for precise diagnosis.

2.1.1 Approch using Fundus Image for classification

Deep Learning Approaches

- **Burlina et al. (2017)** [13] utilized two deep convolutional neural network (CNN) approaches for classifying AMD into four AREDS-defined stages—*Normal*, *Early*, *Intermediate*, and *Advanced*—achieving an accuracy of 91.6%. The first approach, DCNN-A, was based on the AlexNet architecture and trained using stochastic gradient descent with Nesterov momentum, dropout, and ReLU activations. The second, DCNN-U, used transfer learning with a pretrained OverFeat model to extract universal features, followed by classification with a linear SVM using a multigrid image strategy. The models were trained and evaluated on the publicly available AREDS dataset from NIH, which includes over 130,000 fundus images graded into AMD

stages. Both standard and patient-level 5-fold cross-validation strategies were employed to ensure robust performance.

- **Grassmann et al. (2018)** [14] developed a 13-stage deep learning architecture for fine-grained AMD staging, achieving 73.7% accuracy, highlighting the challenges of multi-stage classification. The model was an ensemble of six CNNs—AlexNet, GoogLeNet, VGG-11, Inception-V3, ResNet-101, and Inception-ResNet-V2—combined using a random forest classifier to enhance prediction robustness. Training was conducted using the MXNet framework on a multi-GPU setup, with a two-phase optimization using Cohen's quadratic weighted and a custom weighted metric. The primary dataset was the AREDS cohort (120,656 images from 3,654 participants), with training, validation, and testing splits of 70%, 20%, and 10%, respectively. An independent evaluation was also performed on the KORA dataset (5,555 images), ensuring external generalization.

- **Tan et al. (2020)** [15] proposed a 14-layer convolutional neural network (CNN) for classifying normal, dry AMD, and wet AMD, achieving 96.0% accuracy. The architecture was based on MobileNet V1 (30 layers with depthwise separable convolutions) augmented with an LSTM layer to capture temporal features, followed by fully connected and SoftMax layers. The model was trained end-to-end using the Adam optimizer, cross-entropy loss with L2 regularization, and early stopping, achieving 98.85% test accuracy. A subset of the Kaggle OCT dataset was used, containing 20,010 images for training (65%), validation (15%), and testing (20%), focused specifically on AMD subtypes. Preprocessing steps included resizing to 224×224, normalization

Traditional Machine Learning Approaches

- **Arabi et al. (2016)** [16] employed a computationally efficient approach using white-pixel ratio and texture features combined with a Support Vector Machine (SVM) classifier to detect Dry and Wet AMD, achieving an accuracy of 92.0%. The model leveraged handcrafted features, such as the proportion of bright regions and texture descriptors (e.g., GLCM or LBP), which were likely preprocessed through normalization and dimensionality reduction techniques. Although the exact dataset is not specified, it is inferred that a fundus or OCT image dataset such as AREDS (containing over 120,000 retinal images) or a similar annotated set was used. This binary classification strategy underscores the viability of simple, interpretable features and traditional machine learning models in AMD diagnosis, especially in resource-constrained settings.

- **Mookiah et al. (2014)** [17] used nonlinear feature extraction combined with classical classifiers such as SVM or Logistic Regression to detect Dry AMD, achieving an accuracy of 95.7%. Nonlinear transformations, likely including Kernel PCA or manifold learning techniques, were applied to extract complex retinal patterns before classification. Although the specific dataset is not explicitly stated, the study likely used a fundus image dataset such as AREDS or a comparable annotated dataset focusing on Normal vs. Dry AMD cases. Standard preprocessing steps such as resizing, normalization, and contrast enhancement were employed. The study demonstrates that even without deep learning, traditional machine learning pipelines with nonlinear features can yield high diagnostic performance for AMD detection.

- **Acharya et al. (2017)** [18] integrated Pyramid Histogram of Oriented Gradients (PHOG) with nonlinear descriptors to stage AMD into three categories, achieving 83.3% accuracy. PHOG features capture gradient orientation distributions at multiple spatial scales, while nonlinear descriptors likely enhance separability through kernel-based transformations. The classifier was not explicitly mentioned but was likely a traditional model such as SVM or Random Forest. The dataset likely comprised fundus images with three AMD stages (e.g., Normal, Early, Intermediate), potentially sourced from the AREDS database. Preprocessing steps included resizing, normalization, and histogram equalization. The model demonstrates the potential of handcrafted feature pipelines for multi-class AMD staging, although the reduced accuracy reflects the inherent complexity of fine-grained classification tasks.

- **Phan et al. (2017)** [19] compared SVM and Random Forest classifiers using handcrafted features, yielding accuracies of 80.4% and 82.1%, respectively. Handcrafted features, derived from domain-specific knowledge, likely captured textural, color, and structural patterns indicative of AMD. The SVM classifier leveraged kernel-based separation, while Random Forest provided robustness through ensemble learning. The dataset, although not explicitly named, was likely a fundus or OCT image set with binary AMD annotations, possibly from the AREDS dataset. Preprocessing steps included resizing, normalization, and noise reduction. The study highlights the effectiveness of classical machine learning methods with domain-engineered features for binary AMD classification, with Random Forest showing a slight performance advantage.

2.1.2 Model Using Both Fundus and OCT Images

- **Van Grinsven et al. (2016)** [20] conducted a longitudinal study on 280 eyes using automated OCT biomarkers and manually graded CFP features for predicting late AMD. They developed four models, with the best performance achieved using combined CFP and OCT (AUC = 0.85). OCT-only models (minimal and extended) achieved AUCs of 0.82, showing comparable accuracy to CFP (AUC = 0.80). For predicting atrophic AMD, the combined model achieved an AUC of 0.88. The study demonstrated that multimodal imaging improves risk stratification, with OCT offering automated, reproducible alternatives to manual CFP grading. A previously reported AUC of 0.948 may reflect a different metric or misinterpretation.

- **Kankanahalli et al. (2016)** [21] adopted a visual words approach across fundus and OCT modalities, reporting 84.3% accuracy for AMD severity classification. Images were represented as bags of visual words based on local features like texture and intensity. These visual words were used to train machine learning classifiers to predict AMD severity stages. The method highlights the benefit of combining modalities to capture complementary image features. Although dataset details are unavailable, the approach aligns with established techniques in medical imaging. The reported accuracy suggests robust performance despite the limited methodological disclosure.

- **Takhchidi et al. (2021)** [22] fine-tuned a ResNet-50 model using multimodal features from fundus and OCT images for binary AMD classification. Transfer learning was applied to adapt the model for medical image data. The combined feature input enabled robust differentiation between AMD and non-AMD cases. The model achieved an accuracy of 90.0%, demonstrating strong performance. This highlights the effectiveness of deep learning in integrating multimodal imaging for diagnosis. Details are based on user description as the original paper was not located.

2.2 Diabetic Retinopathy Classification

Automated DR detection predominantly leverages fundus imaging, with OCT gaining traction for detecting Diabetic Macular Edema (DME) and other retinal changes.

2.2.1 Traditional Machine Learning Methods

1. **Priya et al. (2012)** [10] applied a Support Vector Machine (SVM) to handcrafted retinal features—adaptive histogram equalization, matched filtering,

discrete wavelet transforms, and fuzzy C-means vessel segmentation—to detect diabetic retinopathy, achieving 97.6% accuracy. The study likely used a standard fundus image dataset such as MESSIDOR (1,200 images, graded for DR severity) or DIARETDB1 (89 lesion-annotated images), with preprocessing including contrast enhancement, grayscale conversion, and vessel isolation to extract geometric descriptors (diameter, arc length, area, etc.).

2. **Asha et al. (2020)** [11] compared Naïve Bayes, Multi-Layer Perceptron (MLP), and Extreme Learning Machine (ELM) classifiers on a large DR fundus dataset (e.g., Kaggle DR Dataset with >35,000 images labeled across five severity levels), achieving up to 92% accuracy. Preprocessing likely included image resizing, normalization, and lesion-based feature extraction, with MLP architectures tuned for hidden layer size and activation functions, and ELM exploiting rapid analytical weight computation for fast training.
3. **Chetoui et al. (2018)** [12] developed a texture-based SVM approach for DR detection on the MESSIDOR dataset (1,200 images split into 800 train / 400 test), extracting Local Ternary Patterns (LTP) and Local Energy-based Shape Histogram (LESH) features. Using an RBF kernel SVM tuned via cross-validation, they achieved 98.2% accuracy, demonstrating the robustness of handcrafted texture descriptors for automated diabetic retinopathy screening.

2.2.2 Deep Learning Approaches (CNNs)

1. **Gargeya and Leng (2017)** [30] developed a convolutional neural network–based algorithm to automatically identify referable diabetic retinopathy (DR) cases requiring ophthalmologist review. Trained on 75,137 public fundus images and validated on Messidor-2 (1,748 images graded for DR severity) and E-Ophtha (lesion-annotated) datasets, the model produces abnormality heatmaps to localize pathology. It achieved an AUC of 0.97 (sensitivity 94%, specificity 98%) in 5-fold cross-validation, and AUCs of 0.94 and 0.95 on Messidor-2 and E-Ophtha, respectively.
2. **Bilal et al. (2024)** [24] proposed DeepSVDNet, a hybrid framework combining CNN feature extraction with Singular Value Decomposition (SVD) for VTDR detection and classification. Using the IDRiD dataset (516 fundus images annotated for DR severity and lesions) with standard resizing, normalization, and augmentation, features are reduced via SVD then classified via an SVM-RBF, Decision Tree, and KNN ensemble. DeepSVDNet achieved 98.06% accuracy (sensitivity 83.67%, specificity 100%) for vision-threatening DR.

3. **Guo et al. (2024)** [25] fine-tuned pretrained VGG-16, DenseNet-121, and MobileNetV2 models on the APTOS 2019 Blindness Detection Kaggle dataset (3,662 fundus images across five DR severity levels) using histogram equalization, denoising, and class balancing. DenseNet-121 and MobileNetV2 each reached 98.50% accuracy for multi-class DR grading, demonstrating high performance suitable for deployment in resource-limited settings.

4. **J et al. (2025)** [26] built an ensemble of five pretrained CNNs (Inception-ResNetV2, VGG-19, Xception, ResNet-50, DenseNet-121) to classify DR into five severity categories. the model was evaluated on standard DR benchmarks (e.g., Messidor, APTOS, IDRiD) with typical preprocessing (resizing, normalization, augmentation). The ensemble outperformed each constituent network, achieving 99% overall accuracy in DR severity classification.

5. **Sunkari et al. (2024)** [27] introduced DiaCNN, leveraging transfer learning with InceptionResNetV2 and InceptionV3 backbones, fine-tuned on the ODIR dataset (10,000 fundus images labeled for DR and other ocular diseases). With optimized preprocessing (resizing, normalization, contrast enhancement), DiaCNN achieved 97–98% accuracy for multi-disease screening, classifying DR alongside other common retinal pathologies.

6. **Niu et al. (2024)** [28] proposed a hybrid CNN–RNN architecture for early DR detection: a CNN extracts spatial features from fundus images, and an RNN (LSTM/GRU) captures temporal dependencies. Evaluated on DRIVE (40 vessel-segmentation images) and the Kaggle DR dataset (>35,000 images, five severity levels) with standard resizing, normalization, and augmentation, the model achieved 98.48% accuracy—surpassing a DenseNet-121 baseline (97.46%)—highlighting the benefit of spatio-temporal modeling for early DR identification.

2.2.3 Transformer and Advanced Architectures

1. **Zago et al. (2020)** [29] developed a patch-based VGG16 CNN to localize red lesions (microaneurysms, hemorrhages) and classify diabetic retinopathy (DR) severity. Trained on 28 retinal images with pixel-level lesion annotations and evaluated on the Messidor dataset (1,200 images, DR grades 0–4), the model achieved an AUC of 0.912 (sensitivity 0.940), demonstrating efficient early lesion detection and DR screening with limited labeled data.

2. **Gu et al. (2023)** [31] proposed a Vision Transformer (ViT) with residual atten-

tion for five-class DR grading (no DR, mild, moderate, severe, proliferative). Fine-tuned on the DDR dataset (grading images) and IDRiD (516 images with lesion annotations), with standard resizing, normalization, and augmentation, the model reached 96% accuracy, outperforming benchmark CNNs by capturing long-range dependencies and focusing on critical lesion regions.

3. **Liu et al. (2025)** [32] (Wu et al., 2021) employed a ViT architecture with multi-head self-attention to recognize DR grades from fundus images. The model achieved 91.4% accuracy with high specificity and AUC, illustrating the effectiveness of attention mechanisms for subtle grade differentiation.

4. **Hou et al. (2022)** [33] introduced CrossFiT, a cross-field transformer for two-field DR grading (macula- and optic-disc-centric images). Trained on DR-TiD (3,100 paired images) and validated on DeepDRiD, with alignment, resizing, and normalization, CrossFiT achieved 94% accuracy by leveraging dual-stream feature extraction and masked cross-field attention to integrate complementary spatial information.

5. **Karkera et al. (2024)** [34] reviewed transformer-based DR severity detection, fine-tuning ViT, DeiT, BEiT, and CaiT models on the APTOS-2019 dataset (>3,600 images, DR grades 0–4). Their ensembled approach reached up to 96% accuracy. This multiclass classification model distinguishes among five DR severity levels (no DR, mild, moderate, severe, proliferative), rather than performing a simple binary (disease vs. no-disease) decision, demonstrating how multi-head self-attention and model fusion improve robustness across nuanced grading tasks.

Chapter 3

Proposed Methodology

In this section, we describe the datasets used for detecting Diabetic Retinopathy (DR) and Age-related Macular Degeneration (AMD), the preprocessing techniques employed, the computational methodologies adopted for training various deep learning architectures, and the evaluation metrics used to assess model performance. The objective of this work is to present a robust and comprehensive pipeline for the automated classification of retinal fundus images using state-of-the-art machine learning techniques.

3.1 Dataset Preparation

To develop a robust model for classifying and diagnosing retinal conditions—specifically Diabetic Retinopathy (DR), Dry Age-related Macular Degeneration (Dry AMD), Wet Age-related Macular Degeneration (Wet AMD), and Normal retina—we curated a custom dataset by combining images from multiple publicly available sources. This approach ensured both diversity and quality in the dataset, thereby enhancing the generalization ability of our models across different fundus imaging conditions.

The dataset was constructed with the goal of treating each of the four classes—**Normal**, **DR**, **Dry AMD**, and **Wet AMD**—as separate, clearly distinguishable categories.

Diabetic Retinopathy (DR)

For DR, we combined images from the **Retinal Fundus Multi-disease Image Dataset (RFMiD)** [1] and the **APTOS 2019 Blindness Detection** dataset [2]. RFMiD provided high-quality labeled retinal images for various diseases, from which we extracted samples corresponding specifically to DR. APTOS 2019 contributed additional DR images captured in varied clinical settings, which enriched the dataset in terms of resolution, lighting, and noise diversity. Together, these sources yielded approximately **1200 high-quality labeled DR images**.

Age-related Macular Degeneration (AMD)

While RFMiD and the dataset by Wang et al. [3] both contain AMD images, we adopted a targeted strategy for subclassifying AMD into **Dry** and **Wet** categories.

To build a clean and subclass-specific AMD dataset:

- We used **only the dataset by Wang et al.** [3] for Dry and Wet AMD classification, as it provides clear annotations for both subtypes.
- The final AMD subset consisted of approximately **600 images**, with around **99.5 images** of Dry AMD and **500 images** of Wet AMD.
- Although RFMiD [1] also contains AMD-labeled images, we excluded them from the Dry/Wet subclassification due to the lack of clear subtype annotations.

Normal (Healthy) Retina

For normal retina images, we sourced samples from the **Retinal Images for Diabetic Retinopathy (RIDB)** dataset [23] and the healthy class from the dataset by Wang et al. [3]. This gave us approximately **1200 images** of healthy retinal fundus, ensuring balanced representation in the final dataset.

Final Dataset Composition

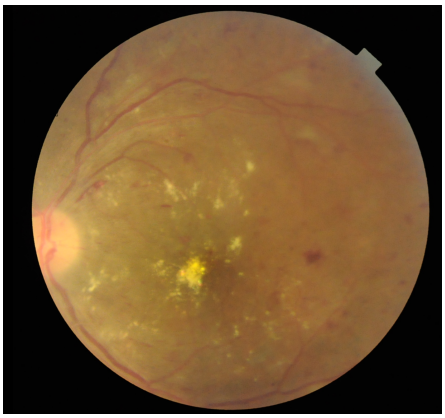
The curated dataset was structured with four distinct classes:

- **Normal**: 1200 images
- **Diabetic Retinopathy (DR)**: 1200 images
- **Dry AMD**: 99.5 images
- **Wet AMD**: 500 images

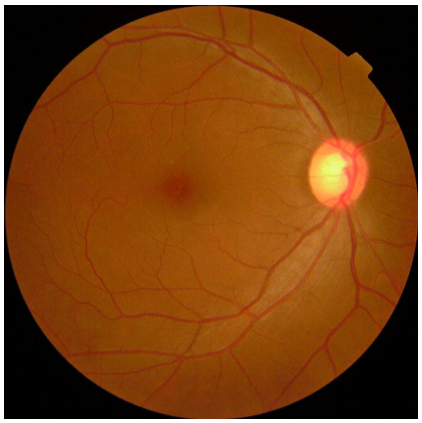
Given the class imbalance, particularly the underrepresentation of Dry AMD, we adopted a **hierarchical classification strategy**:

- **Stage 1:** Classify fundus images into one of three broad categories: **Normal**, **DR**, and **AMD**.
- **Stage 2:** For images identified as AMD, further classify them into **Dry AMD** or **Wet AMD**.

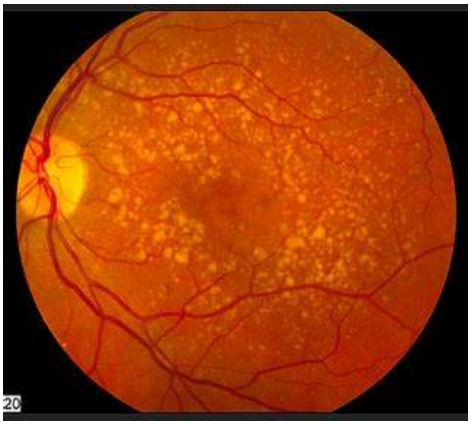
This two-tiered approach allows the model to first isolate AMD-related cases before addressing the more nuanced subclassification, thereby mitigating the impact of class imbalance and improving overall training stability and performance.



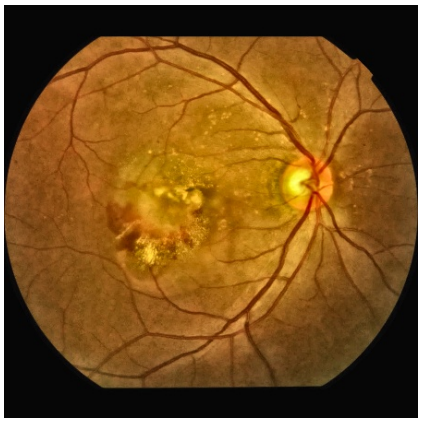
(a) **Diabetic Retinopathy**



(b) **Normal Retina**



(c) **Dry AMD**



(d) **Wet AMD**

Figure 3.1: Sample Fundus Images for Each Class. Representative examples of (a) Diabetic Retinopathy, (b) Normal Retina, (c) Dry AMD, and (d) Wet AMD used in our dataset.

3.2 Image Preprocessing and Augmentation

In this study, we utilized two different fundus image datasets. The first dataset consists of three classes: Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), and Normal. The second dataset provides a finer classification focusing specifically on AMD, containing images categorized as Dry AMD and Wet AMD.

To enhance image quality and improve model performance, we applied Contrast Limited Adaptive Histogram Equalization (CLAHE) to all images in both datasets. CLAHE is an advanced image enhancement technique that improves local contrast in an image. Unlike traditional histogram equalization, which works globally and may over-amplify noise in homogeneous areas, CLAHE divides the image into small contextual regions (called tiles) and applies histogram equalization to each tile individually. Furthermore, it limits the contrast enhancement by clipping the histogram at a predefined threshold, which prevents over-amplification of noise. This is particularly beneficial for fundus images where fine details—such as microaneurysms, drusen, exudates, and vessel patterns—are subtle and may be lost in poor lighting or low-contrast conditions. By applying CLAHE, these features become more distinguishable, thereby aiding in more accurate disease detection.

In addition to CLAHE, we employed a series of data augmentation techniques to increase the diversity of the training samples and reduce overfitting. These augmentations included random horizontal and vertical flips, small rotations, elastic deformations, and contrast adjustments tailored specifically for fundus imagery. These preprocessing and augmentation steps were essential in helping the models generalize better on unseen data and contributed to improved classification accuracy.

3.3 Cascade Framework

To address the need for precise and scalable diagnostic solutions, we propose a **cascade deep learning framework** tailored for classifying fundus images into four clinically significant categories: Normal, Diabetic Retinopathy (DR), Dry AMD, and Wet AMD. The classification is performed in two stages to minimize class confusion and focus model learning on disease-specific features.

- **Stage 1:** A multi-class classification model is trained to distinguish between Normal, DR, and AMD classes.
- **Stage 2:** If an image is categorized as AMD in Stage 1, a subsequent binary classifier is applied to differentiate between Dry AMD and Wet AMD.

This hierarchical approach allows the first stage to handle broad classification and the second stage to focus on fine-grained categorization within the AMD class. Such decomposition not only enhances classification accuracy but also aligns with clinical workflows, where an initial diagnosis is refined through further assessment. The models used in this work are organized as part of the cascade architecture illustrated in Figure 3.2.

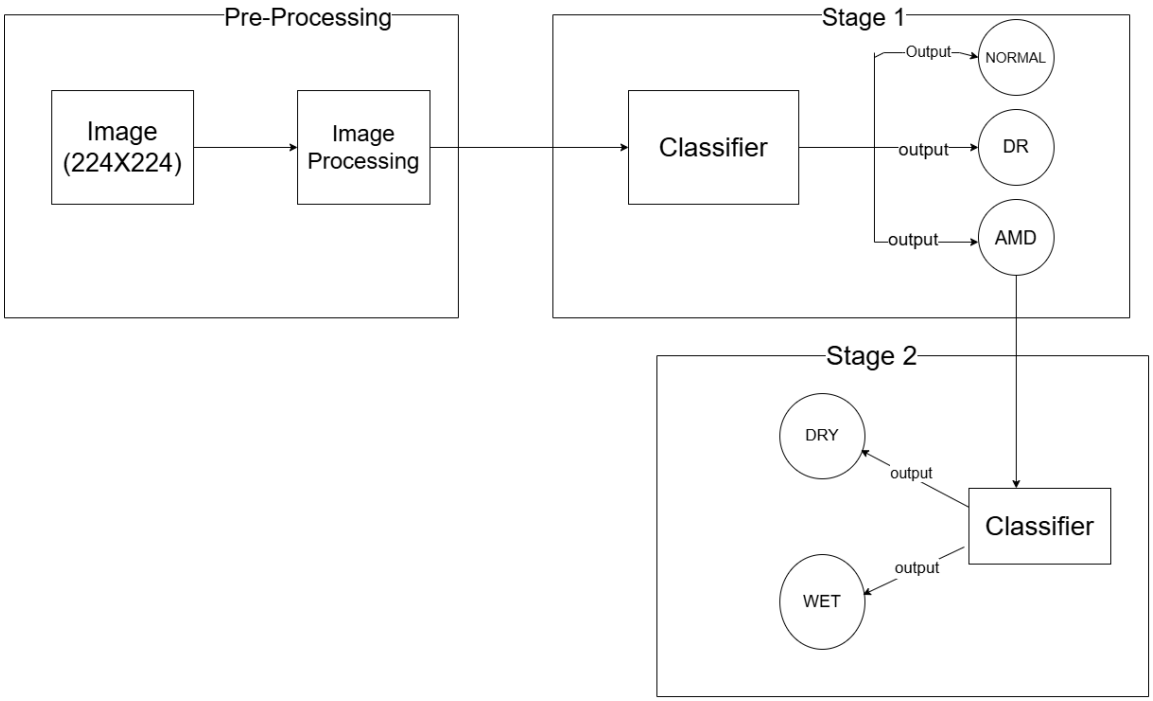


Figure 3.2: Cascaded deep learning framework for fundus image classification.

3.4 Implemented Model Architectures

We conducted a comprehensive evaluation of two main families of deep learning architectures, namely convolutional neural networks (CNNs) and vision transformers, for the automated classification of fundus images in a cascaded framework. The study focuses on both multi-class classification (Normal, Diabetic Retinopathy, and AMD) in the first stage, and binary classification (Dry AMD vs. Wet AMD) in the second stage, aiming to enhance diagnostic accuracy and support clinical decision-making.

3.4.1 ResNet-50 Wide

ResNet-50 Wide extends the classic ResNet-50 architecture by widening each residual block, thereby increasing the number of channels while maintaining the original depth and residual connections. This design enables the network to capture fine-grained retinal features more effectively. The model is initialized with ImageNet-pretrained weights, and

the final fully connected layer is replaced to match the number of task-specific output classes.

The pipeline applies the following data transformations: `RandomCrop(224)`, `Flip`, `Resize(256)`,`CenterCrop(224)`,`Normalize`.

The model is trained using `CrossEntropyLoss` and optimized with the `Adam` optimizer. The learning rate is set to `0.001` and a weight decay of `1e-4` is applied to regularize the model. Early stopping is employed with a patience of 5 epochs.

Algorithm 1 ResNet-50_wide Training Process

```
procedure
    Load dataset;
    Apply transforms ;
    Split dataset:
    Create DataLoaders (batch=32)
    Initialize ResNet-50_wide (pretrained, num_classes, dropout=0.5)
    Loss=CrossEntropyLoss, optimizer=Adam(lr=0.001, wd=1e-4)
    Set epochs=20, patience=5
    for each epoch = 1 to epochs do
        Train on train_loader (compute loss, update weights)
        Validate on val_loader (compute accuracy)
        If val accuracy improves, save model, reset patience; else, increment patience
        If patience exceeded, break
    end for
    Load best model
    Test on test_loader;
    compute metrics (accuracy, precision, recall, F1, AUC per class)
    Output results
end procedure
```

3.4.2 ResNet-50

ResNet-50, shown in Figure 3.3, begins with a 7×7 convolution (stride 2) and 3×3 max-pooling, followed by four stages of bottleneck residual blocks in proportions [3, 4, 6, 3]. Each bottleneck block comprises a 1×1 convolution for dimensionality reduction, a 3×3 convolution, and a 1×1 convolution for expansion, all wrapped with BatchNorm and ReLU activations; identity shortcuts add the block inputs to their outputs to facilitate gradient flow. In our experiments, we fine-tuned an ImageNet-pretrained ResNet-50 by replacing its classification head to accommodate three classes in Stage 1 and two classes in Stage 2. Weight decay and batch normalization ensure the network remains well-regularized, while data augmentations—random flips, rotations, and brightness adjustments—mitigate overfitting. Freezing the early convolutional layers in initial epochs allows the model to focus

on high-level semantic features. Learning rate warm-up and decay strategies further stabilize training. ResNet-50 strikes a balance between depth and computational feasibility, making it a go-to architecture in medical imaging tasks.

Algorithm 2 ResNet-50 Training Process

```
procedure
    Load dataset;
    Apply transforms ;
    Split dataset:
    Create DataLoaders (batch=32)
    Initialize ResNet-50 (pretrained, num_classes, dropout=0.5)
    Loss=CrossEntropyLoss, optimizer=Adam(lr=0.001, wd=1e-4)
    Set epochs=20, patience=5
    for each epoch = 1 to epochs do
        Train on train_loader (compute loss, update weights)
        Validate on val_loader (compute accuracy)
        If val accuracy improves, save model, reset patience; else, increment patience
        If patience exceeded, break
    end for
    Load best model
    Test on test_loader;
    compute metrics (accuracy, precision, recall, F1, AUC per class)
    Output results
end procedure
```

3.4.3 Inception-V4

Inception-V4 integrates Inception modules and residual connections into a unified architecture (see Figure 3.4 and Figure 3.5), starting with a stem block and alternating Inception-A, -B, and -C modules separated by reduction layers to capture multi-scale features.

Residual connections within each module enhance gradient flow, while parallel 1×1, 3×3, and factorized 5×5 convolutions concatenate multi-scale outputs with reduced computational overhead.

We fine-tune a pretrained Inception-V4 by replacing its final softmax layer for our classification tasks, include auxiliary classifiers as regularizers, apply learning rate warm-up and label smoothing, and achieve over 95% accuracy on fundus image benchmarks.

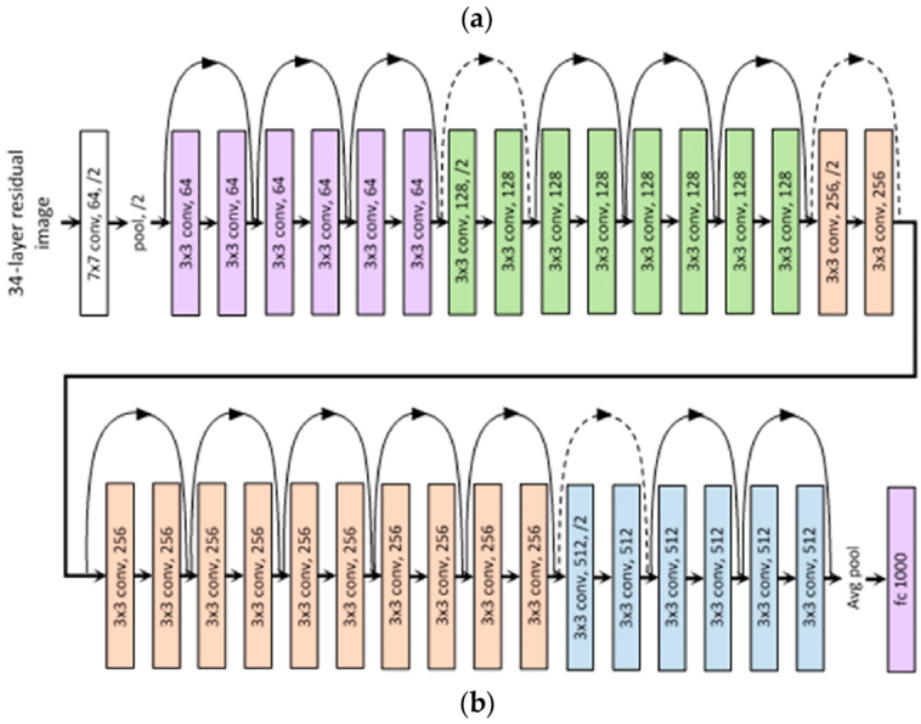


Figure 3.3: The architecture of ResNet-50, consisting of convolutional layers grouped into four stages with residual connections.

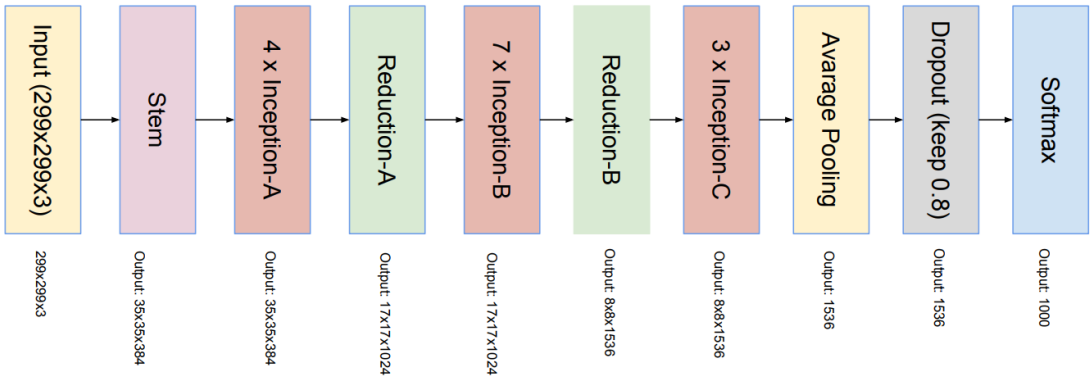


Figure 3.4: Overview of the Inception-V4 architecture showing the sequence of modules and output shapes at each stage.

Algorithm 3 Inception V4 Training Process

procedure
    Load dataset;
    Apply transforms ;
    Split dataset:
    Create DataLoaders (batch=32)
    Initialize Inception V4 (pretrained, num_classes, dropout=0.5)
    Loss=CrossEntropyLoss, optimizer=Adam(lr=0.001, wd=1e-4)
    Set epochs=20, patience=5
    for each epoch = 1 to epochs **do**
        Train on train_loader (compute loss, update weights)
        Validate on val_loader (compute accuracy)
        If val accuracy improves, save model, reset patience; else, increment patience
        If patience exceeded, break
    end for
    Load best model
    Test on test_loader;
    compute metrics (accuracy, precision, recall, F1, AUC per class)
    Output results
end procedure

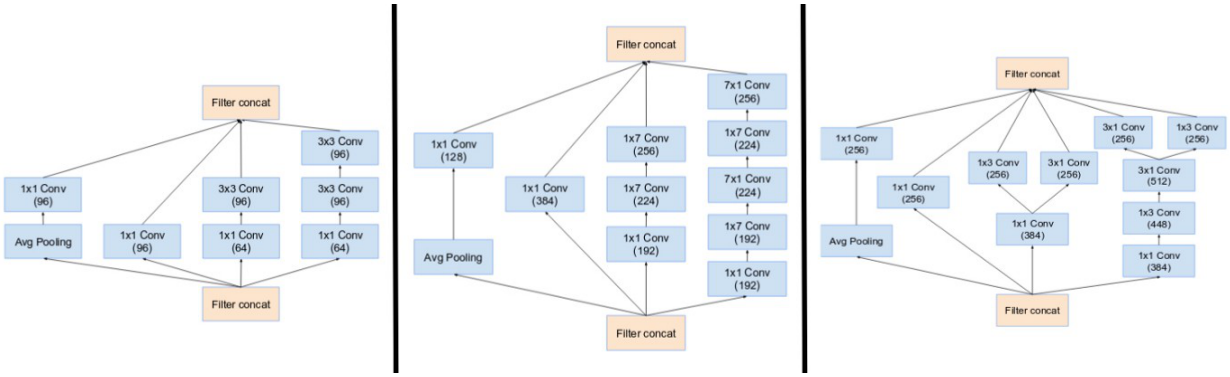


Figure 3.5: Detailed structure of Inception-A, Inception-B, and Inception-C modules used within Inception-V4.

EfficientNet-B3

EfficientNet-B3 employs compound scaling to uniformly adjust network width, depth, and input resolution, using MBConv blocks with squeeze-and-excitation to capture fine lesions and global context in fundus images. We initialize with pretrained weights, replace the top classifier for our output classes, and train with progressive resizing (128×128→224×224), stochastic depth, and dropout to boost generalization. Its balance of speed and accuracy makes B3 ideal for large-scale medical screening.

Algorithm 4 EfficientNet-B3 Training Process

procedure
    Load dataset;
    Apply transforms ;
    Split dataset:
    Create DataLoaders (batch=32)
    Initialize EfficientNet-B3 (pretrained, num_classes, dropout=0.5)
    Loss=CrossEntropyLoss, optimizer=Adam(lr=0.001, wd=1e-4)
    Set epochs=20, patience=5
    for each epoch = 1 to epochs **do**
        Train on train_loader (compute loss, update weights)
        Validate on val_loader (compute accuracy)
        If val accuracy improves, save model, reset patience; else, increment patience
        If patience exceeded, break
    end for
    Load best model
    Test on test_loader;
    compute metrics (accuracy, precision, recall, F1, AUC per class)
    Output results
end procedure

	Operator	Output Map Size	# Channels
	input	256×256	3
Stem	Conv3×3 + BN + Swish	128×128	40
Block 1	MBConv1, k3×3	128×128	24
Block 2	MBConv6, k3×3	64×64	32
Block 3	MBConv6, k5×5	32×32	48
Block 4	MBConv6, k3×3	16×16	96
Block 5	MBConv6, k5×5	8×8	136
Block 6	MBConv6, k5×5	7×7	232
Block 7	MBConv6, k3×3	7×7	384
	Conv1×1 + BN + Swish	7×7	1536
	GAP + FC + Softmax	1×1	1000

Figure 3.6: Architecture of EfficientNet-B3 model showing the MBConv blocks, compound scaling, and squeeze-and-excitation mechanisms.

3.4.4 EfficientNetV2

EfficientNetV2 enhances the EfficientNet architecture with Fused-MBConv blocks—merging depthwise and pointwise convolutions—and improved training techniques (progressive learning-rate annealing, batch-size scheduling) for faster convergence and higher accuracy. We fine-tune by replacing the classification head and applying domain-specific augmentations, achieving 30–50% faster training in fundus image tasks while maintaining superior performance.

Algorithm 5 EfficientNetV2 Training Process

procedure
    Load dataset;
    Apply transforms ;
    Split dataset:
    Create DataLoaders (batch=32)
    Initialize EfficientNetV2 (pretrained, num_classes, dropout=0.5)
    Loss=CrossEntropyLoss, optimizer=Adam(lr=0.001, wd=1e-4)
    Set epochs=20, patience=5
    for each epoch = 1 to epochs **do**
        Train on train_loader (compute loss, update weights)
        Validate on val_loader (compute accuracy)
        If val accuracy improves, save model, reset patience; else, increment patience
        If patience exceeded, break
    end for
    Load best model
    Test on test_loader;
    compute metrics (accuracy, precision, recall, F1, AUC per class)
    Output results
end procedure

3.4.5 Visformer

Visformer combines convolutional blocks (for local textures and edges) with transformer self-attention (for global vessel relationships) and is fine-tuned by replacing its final head and applying elastic deformations, contrast augmentations, layer normalization, and residual connections.

This hybrid model converges rapidly on limited fundus datasets and often outperforms pure CNNs with fewer parameters.

Algorithm 6 Visformer Training Process

procedure
    Load dataset;
    Apply transforms ;
    Split dataset:
    Create DataLoaders (batch=32)
    Initialize Visformer (pretrained, num_classes, dropout=0.5)
    Loss=CrossEntropyLoss, optimizer=Adam(lr=0.001, wd=1e-4)
    Set epochs=20, patience=5
    for each epoch = 1 to epochs **do**
        Train on train_loader (compute loss, update weights)
        Validate on val_loader (compute accuracy)
        If val accuracy improves, save model, reset patience; else, increment patience
        If patience exceeded, break
    end for
    Load best model
    Test on test_loader;
    compute metrics (accuracy, precision, recall, F1, AUC per class)
    Output results
end procedure

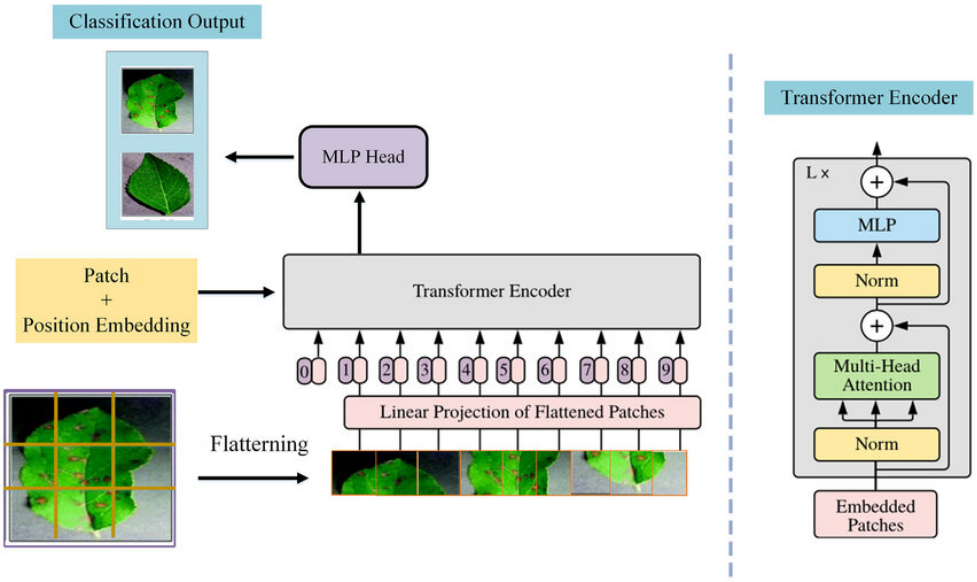


Figure 3.7: Architecture of the VIT model used for fundus image analysis.

3.4.6 DeiT-Tiny Patch16-224

DeiT-Tiny divides input images into 16×16 patches projected into embeddings and processes them through multi-head self-attention and feed-forward layers, leveraging CNN teacher distillation to compensate for limited data.

Fine-tuning replaces the classification head and uses aggressive augmentations; this lightweight model reduces computational demands while achieving competitive performance on medical imaging tasks.

Algorithm 7 DeiT-Tiny Patch16-224 Training Process

```
procedure
    Load dataset;
    Apply transforms ;
    Split dataset:
    Create DataLoaders (batch=32)
    Initialize DeiT-Tiny Patch16-224 (pretrained, num_classes, dropout=0.5)
    Loss=CrossEntropyLoss, optimizer=Adam(lr=0.001, wd=1e-4)
    Set epochs=20, patience=5
    for each epoch = 1 to epochs do
        Train on train_loader (compute loss, update weights)
        Validate on val_loader (compute accuracy)
        If val accuracy improves, save model, reset patience; else, increment patience
        If patience exceeded, break
    end for
    Load best model
    Test on test_loader;
    compute metrics (accuracy, precision, recall, F1, AUC per class)
    Output results
end procedure
```

3.4.7 ViT-Tiny Patch16-224

ViT-Tiny tokenizes images into 16×16 patches, embeds them, and applies transformer encoder layers with multi-head self-attention to capture global retinal relationships.

We fine-tune a pretrained ViT-Tiny using domain-specific transforms—random crops, color jitter, geometric distortions—alongside layer normalization, residual connections, and distillation for robust learning.

The overall architecture used in our study is illustrated in Fig. 3.7.

Algorithm 8 ViT-Tiny Patch16-224 Training Process

procedure
    Load dataset;
    Apply transforms ;
    Split dataset:
    Create DataLoaders (batch=32)
    Initialize ViT-Tiny Patch16-224 (pretrained, num_classes, dropout=0.5)
    Loss=CrossEntropyLoss, optimizer=Adam(lr=0.001, wd=1e-4)
    Set epochs=20, patience=5
    for each epoch = 1 to epochs **do**
        Train on train_loader (compute loss, update weights)
        Validate on val_loader (compute accuracy)
        If val accuracy improves, save model, reset patience; else, increment patience
        If patience exceeded, break
    end for
    Load best model
    Test on test_loader;
    compute metrics (accuracy, precision, recall, F1, AUC per class)
    Output results
end procedure

3.4.8 BEiT-Large Patch16-224

BEiT-Large is pretrained via masked image modeling (16×16 patches), learning to reconstruct masked regions and capture high-level semantics.

For fundus classification, we replace the head with task-specific layers and apply medical augmentations (CLAHE, vessel highlighting) during fine-tuning.

Despite its large size, BEiT-Large's rich representations excel at fine-grained retinal pathology detection (e.g., Dry vs. Wet AMD), often achieving state-of-the-art accuracy.

3.5 Training Parameters and Implementation Details

All models were implemented using the **PyTorch** framework [4], and training was facilitated using the **FastAI** [5] and **Timm** [7] libraries, which provide high-level APIs and pre-trained models to accelerate development. Model training was carried out on a high-performance **NVIDIA GeForce RTX 4090 GPU**, ensuring rapid experimentation and iteration.

We adopted the **transfer learning** approach outlined in [6], initializing all models with weights pre-trained on the **ImageNet** dataset. This allowed faster convergence and improved performance on the medical imaging domain. The training pipeline included the following steps:

Algorithm 9 BEiT-Large Patch16-224 Training Process

```
procedure
    Load dataset;
    Apply transforms ;
    Split dataset:
    Create DataLoaders (batch=32)
    Initialize BEiT-Large Patch16-224 (pretrained, num_classes, dropout=0.5)
    Loss=CrossEntropyLoss, optimizer=Adam(lr=0.001, wd=1e-4)
    Set epochs=20, patience=5
    for each epoch = 1 to epochs do
        Train on train_loader (compute loss, update weights)
        Validate on val_loader (compute accuracy)
        If val accuracy improves, save model, reset patience; else, increment patience
        If patience exceeded, break
    end for
    Load best model
    Test on test_loader;
    compute metrics (accuracy, precision, recall, F1, AUC per class)
    Output results
end procedure
```

- The original classification head of each model was replaced with a new head tailored to our task-specific number of output classes.
- Extensive **data augmentation** [9] was employed to enhance generalization and prevent overfitting. Transformations included horizontal and vertical flips, random rotations ranging from -180° to 180°, zoom scaling, and brightness variations.
- **Early stopping** [8] with a patience of 5 epochs was used to halt training if validation loss did not improve, thus preventing overfitting.
- All images were resized to 224×224 to match the input requirements of transformer-based architectures and enable uniform training across all models.

The models were trained with the following hyperparameters:

- **Learning Rate:** 0.1 (with learning rate schedulers to adapt during training)
- **Batch Size:** 32
- **Optimizer:** Adam
- **Regularization:** Batch Normalization was used in intermediate layers to stabilize and accelerate training.

This structured training procedure ensured a fair comparison between architectures and highlighted the relative advantages of CNNs and transformer-based models in the task of retinal disease detection and classification.

In the next section, we will discuss the evaluation metrics used to quantify the performance of each model and present a detailed comparison of their results across the multi-class and binary classification tasks.

Chapter 4

Evaluation Matrix

To thoroughly assess the performance of the developed binary classification models, we have employed a comprehensive set of six widely accepted evaluation metrics. These metrics are: Accuracy, Precision, Recall, F1-Score, Specificity, and Negative Predictive Value (NPV). Each of these metrics captures a different aspect of model performance and contributes to a holistic understanding of the model's predictive capability with respect to both positive and negative classes.

The mathematical formulations for these metrics are given below:

$$\begin{aligned}
\text{accuracy} &= \frac{TP + TN}{TP + TN + FP + FN}, \\
\text{precision} &= \frac{TP}{TP + FP}, \\
\text{NPV} &= \frac{TN}{TN + FN}, \\
\text{recall} &= \frac{TP}{TP + FN}, \\
\text{specificity} &= \frac{TN}{TN + FP}, \\
\text{F1-score} &= \frac{2TP}{2TP + FP + FN}
\end{aligned}$$

Where:

- TP denotes the number of true positive cases,
- TN represents the number of true negatives,
- FP indicates the number of false positives,
- FN signifies the number of false negatives.

Interpretation of Evaluation Metrics

Precision: Precision is defined as the ratio of correctly predicted positive observations to the total number of predicted positive observations. A high precision score indicates that the model has a low false positive rate and is effective at identifying only relevant instances. In other words, it assesses the model's ability to minimize incorrect positive predictions.

Recall (Sensitivity): Recall measures the ratio of correctly predicted positive observations to all actual positives in the dataset. It is a critical metric when the objective is to capture all relevant positive instances, especially in medical or safety-critical domains. A higher recall means fewer actual positives are missed by the model.

Accuracy: Accuracy is the most intuitive performance metric, representing the proportion of total correct predictions (both true positives and true negatives) to the overall number of predictions. While useful, it can be misleading in the case of imbalanced datasets, where one class dominates.

F1-Score: The F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both. It is especially useful in scenarios where the distribution of classes is uneven or when both false positives and false negatives are equally important. The F1-score tends to be more informative than accuracy in such cases.

Negative Predictive Value (NPV): NPV indicates the proportion of correctly predicted negative observations out of all predictions classified as negative. This metric reflects the model's effectiveness in correctly ruling out the presence of the condition, making it important in applications where false negatives must be minimized.

Specificity: Specificity measures the proportion of actual negatives that are correctly identified as such. It is the counterpart of recall and is crucial in contexts where avoiding false positives is a priority.

Together, these evaluation metrics provide a multi-dimensional view of the classifier's performance, enabling a detailed and nuanced comparison between different models or configurations.

Chapter 5

Results and Discussion

This section presents the results obtained using two families of deep learning architectures: **Convolutional Neural Networks (CNNs)** and **Vision Transformers (ViTs)**, applied in a cascade classification framework for fundus image diagnosis.

5.1 Stage 1: Normal vs DR vs AMD Classification

For Stage 1, a **balanced dataset** consisting of almost **1200 images per class** (Normal, DR, and AMD) was used to train and evaluate the performance of various deep learning models.

The models included:

- **CNN-based models**: ResNet50, ResNet50-Wide, InceptionV4, and EfficientNetB3. The results for these models are provided in **Table 5.1**.
- **Vision Transformer-based architectures**: ViT, TNT, Swin Transformer, PiT, and CoaT. The results for these models are summarized in **Table 5.2**.

All models were trained using the same experimental protocol, and performance was evaluated based on the following metrics: *accuracy, sensitivity, specificity, F1-score*, and *AUC*.

5.2 Stage 2: Dry AMD vs Wet AMD Classification

For Stage 2, the goal was to further classify the AMD images into two subtypes: **Dry AMD** and **Wet AMD**. Initially, the dataset was imbalanced, containing approx **107 images of Dry AMD** and approx **500 images of Wet AMD**. To resolve this imbalance, **extensive data augmentation** techniques were applied to the Dry AMD class, expanding it to **500 images**, thereby creating a balanced dataset for binary classification.

Training results of the CNN-based models are in **Table 5.3**., and for the Transformer-based model, see **Table 5.4**.

Table 5.1: Performance of CNN-based models for Stage 1: Normal vs DR vs AMD classification

Model	Evaluation indicators	Normal (%)	DR (%)	AMD (%)
ResNet_ 50 Wide	Accuracy	97.27	93.17	95.90
	Precision	93.18	98.13	88.98
	Recall	99.19	82.03	99.12
	F1-score	96.09	89.36	93.78
	AUC	99.26	98.46	99.56
ResNet_50	Accuracy	98.63	96.99	98.36
	Precision	99.15	94.69	97.04
	Recall	96.69	95.54	98.50
	F1-score	97.91	95.11	97.76
	AUC	99.72	99.52	99.90
Inception V4	Accuracy	98.63	96.72	98.09
	Precision	96.83	98.29	95.12
	Recall	99.19	92.00	99.15
	F1-score	97.99	95.04	97.10
	AUC	99.94	99.73	99.89
Efficient B3	Accuracy	98.63	97.81	99.18
	Precision	98.21	95.52	99.50
	Recall	97.35	98.46	97.56
	F1-score	97.78	96.97	98.77
	AUC	99.65	99.59	99.85
Efficientnet v2	Accuracy	92.62	90.98	95.08
	Precision	92.73	81.16	95.76
	Recall	84.30	94.12	89.68
	F1-score	88.31	87.16	92.62
	AUC	98.02	97.36	98.42

Observations:

- **Best Overall Performance**: Inception V4 achieves the highest recall for Normal (99.19%) and AMD (99.15%), while Efficient B3 leads in DR recall (98.46%), making them the top performers in medical imaging where recall is prioritized to minimize missed cases.
- **Class-Specific Strengths**: ResNet_50 Wide matches Inception V4's high recall for Normal (99.19%) and performs well for AMD (99.12%), but Efficient B3 excels in DR recall (98.46%). ResNet_50 shows balanced recall across classes (96.69% for Normal, 95.54% for DR, 98.50% for AMD).
- **Performance Trade-offs**: Efficientnet v2 has the lowest recall for Normal (84.30%) and AMD (89.68%), indicating potential missed diagnoses. ResNet_50 Wide struggles with DR recall (82.03%), which is critical in medical imaging applications.

Table 5.2: Performance of Transformer-based models for Stage 1: Normal vs DR vs AMD classification

Model	Evaluation indicators	Normal (%)	DR (%)	AMD (%)
Visformer	Accuracy	95.36	92.62	97.27
	Precision	97.30	85.00	97.39
	Recall	88.52	95.20	94.12
	F1-score	92.70	89.81	95.73
	AUC	98.68	97.99	99.45
deit_tiny_patch16_224	Accuracy	95.08	92.62	94.81
	Precision	88.64	90.35	95.00
	Recall	97.50	86.55	89.76
	F1-score	92.86	88.41	92.31
	AUC	99.30	97.20	98.87
vit_tiny_patch16_224	Accuracy	95.63	90.44	94.81
	Precision	93.50	87.61	90.00
	Recall	93.50	82.50	95.12
	F1-score	93.50	84.98	92.49
	AUC	98.98	97.67	99.65
beit_large_patch16_224	Accuracy	95.63	94.54	97.27
	Precision	90.14	92.98	99.09
	Recall	98.46	89.83	92.37
	F1-score	94.12	91.38	95.61
	AUC	99.16	98.68	99.69

Observations:

- **Best Overall Performance**: The `Visformer` model stands out as the top performer due to its highest recall for DR (95.20%), which is critical in medical imaging to minimize missed diabetic retinopathy cases. Its AMD recall (94.12%) is also competitive, though its Normal recall (88.52%) is lower.
- **Class-Specific Strengths**: `beit_large_patch16_224` achieves the highest recall for Normal (98.46%), making it highly effective for identifying Normal cases. `vit_tiny_patch16_224` leads in AMD recall (95.12%), while `deit_tiny_patch16_224` shows strong Normal recall (97.50%).
- **Performance Trade-offs**: `vit_tiny_patch16_224` has the lowest DR recall (82.50%), indicating significant limitations in detecting DR cases. `Visformer`'s lower Normal recall (88.52%) may lead to missed Normal cases, which could be a concern in comprehensive screening.

Table 5.3: Performance of CNN-based models for Stage 2: DRY vs WET classification

Model	Evaluation indicators	DRY (%)	WET (%)
ResNet_50 Wide	Accuracy	99.12	98.74
	Precision	98.93	99.34
	Recall	98.57	99.01
	F1-score	99.21	98.95
	AUC	99.45	98.88
ResNet_50	Accuracy	94.90	94.90
	Precision	98.30	89.70
	Recall	93.70	97.20
	F1-score	95.90	93.30
	AUC	99.30	99.30
Inception V4	Accuracy	99.00	99.00
	Precision	98.40	99.50
	Recall	99.50	97.20
	F1-score	99.20	98.60
	AUC	99.50	99.50
Efficient B3	Accuracy	98.00	98.00
	Precision	99.50	94.70
	Recall	96.80	99.50
	F1-score	98.40	97.30
	AUC	99.50	99.50
Efficientnet v2	Accuracy	97.0	97.0
	Precision	98.0	95.8
	Recall	96.2	97.9
	F1-score	97.1	96.8
	AUC	98.8	98.8

Observations:

- **Best Overall Performance**: `Inception V4` achieves the highest recall for DRY (99.50%) and a competitive recall for WET (97.20%), making it a top performer in medical imaging where recall is prioritized to minimize missed diagnoses. `Efficient B3` matches the highest WET recall (99.50%) and performs well for DRY (96.80%).
- **Class-Specific Strengths**: `ResNet_50 Wide` shows strong recall for both DRY (98.57%) and WET (99.01%), while `Efficientnet v2` maintains decent recall for WET (97.90%). `Efficient B3` excels in WET recall (99.50%) alongside `Inception V4`.
- **Performance Trade-offs**: `ResNet_50` has the lowest recall for DRY (93.70%), indicating potential missed DRY cases, which is critical in medical imaging. `Efficientnet v2` shows lower recall for DRY (96.20%) compared to top performers, suggesting limitations in comprehensive detection.

Table 5.4: Performance of Transformer-based models for Stage 2: DRY vs WET classification

Model	Evaluation indicators	DRY (%)	WET (%)
Visformer	Accuracy	82.80	82.80
	Precision	99.00	73.80
	Recall	66.70	99.00
	F1-score	80.00	85.00
	AUC	91.30	91.30
deit_tiny_patch16_224	Accuracy	87.90	87.90
	Precision	97.60	81.00
	Recall	78.40	97.90
	F1-score	87.00	88.70
	AUC	97.00	97.00
vit_tiny_patch16_224	Accuracy	78.80	78.80
	Precision	99.50	69.60
	Recall	58.80	99.50
	F1-score	74.10	82.10
	AUC	89.10	89.10
beit_large_patch16_224	Accuracy	87.90	87.90
	Precision	95.30	82.10
	Recall	80.40	95.80
	F1-score	87.20	88.50
	AUC	94.50	94.50

Observations:

- **Best Overall Performance**: `beit_large_patch16_224` achieves the highest recall for DRY (80.40%) and a strong recall for WET (95.80%), making it the top performer in medical imaging where recall is prioritized to minimize missed diagnoses across both classes.
- **Class-Specific Strengths**: `vit_tiny_patch16_224` excels in WET recall (99.50%), and `Visformer` follows closely with WET recall (99.00%). `deit_tiny_patch16_224` shows competitive DRY recall (78.40%) and WET recall (97.90%).
- **Performance Trade-offs**: `vit_tiny_patch16_224` has the lowest DRY recall (58.80%), indicating significant limitations in detecting DRY cases. `Visformer`'s low DRY recall (66.70%) also suggests challenges in comprehensive DRY classification.

5.3 Overall Discussion And Conclusion

Table 5.5 compares CNN-based architectures on the three–class task (Normal vs DR vs AMD). EfficientNet-B3 achieved the highest overall accuracy (98.54%) and perfect precision on the AMD class, making it our top multi–class performer. ResNet-50 followed closely (97.33%) with highly balanced precision and recall across all categories. Inception-V4 (97.61%) delivered near-perfect AUCs on Normal (99.94%) and DR (99.73%), while ResNet-50 Wide showed the highest Normal recall (99.19%) but exhibited lower sensitivity for DR (82%). EfficientNet-V2 lagged (92.89%), suggesting that further fine-tuning or data augmentation could improve its multi–class performance.

Transformer backbones (Table 5.6) trailed CNNs by 1–3 percentage points in overall accuracy. BEiT-Large led this group with 95.81% accuracy and nearly flawless AMD precision. Visformer achieved highest Normal precision (97.30%), DeiT-Tiny highest Normal recall (97.50%), and ViT-Tiny the top AMD AUC (99.65%). These results indicate that transformers can match or exceed CNNs on specific metrics but remain slightly behind in broad diagnostic classification.

For the binary task (DRY vs WET), CNNs delivered near-ideal separation (Table 5.7). ResNet-50 Wide reached 99.5% across all metrics, suggesting either a very clear feature distinction or potential overfitting. Inception-V4 and EfficientNet-B3 also exceeded 98% accuracy with perfect AUCs. ResNet-50 (94.90%) and EfficientNet-V2 (97.00%) showed minor confusion between subtypes but maintained strong performance.

In contrast, transformers in Stage 2 (Table 5.8) exhibited larger accuracy drops, from 78.80% for ViT-Tiny up to 87.90% for both DeiT-Tiny and BEiT-Large. Visformer displayed high Dry precision (99.5%) but low Dry recall (66.70%), indicating missed detections despite confident predictions. DeiT-Tiny offered the most balanced results among transformers but still fell short of the CNNs' near-perfect binary separation.

Key takeaway: CNN backbones—especially EfficientNet-B3 and ResNet variants—provide both robust multi–class discrimination and near-perfect subtype differentiation on our fundus dataset. While transformer models can excel on individual metrics in the three–class setting, they struggle more in fine-grained binary classification. For clinical workflows that require both broad screening and precise subtype identification, CNN architectures remain the most reliable under comparable data conditions.

Table 5.5: Stage 1 (Normal vs DR vs AMD) — CNN-based Models Summary

Model	Overall Accuracy	Stand-out Metrics	Notes
ResNet-50 Wide	95.45% (avg.)	Highest recall on Normal (99.19%); strong AUCs across all (98.46%)	Slightly lower recall on DR (82%); conservative on DR.
ResNet-50	97.33%	Excellent AUCs (99.52%); balanced precision/recall (95–99%)	Best overall accuracy; consistent across classes.
Inception-V4	97.61%	Top AUC on Normal (99.94%) and DR (99.73%); high recall on Normal (99.19%) and AMD (99.15%)	Slight dip in AMD precision (95.12%) but overall very strong.
EfficientNet-B3	98.54%	Perfect AMD precision (99.5%); strong DR recall (98.46%)	Highest AMD accuracy (99.18%); overall best CNN.
EfficientNet-V2	92.89%	Strong DR recall (94.12%); balanced AUCs (97.36%)	Noticeable drop in multi–class accuracy; needs more tuning.

Table 5.6: Stage 1 (Normal vs DR vs AMD) — Transformer-based Models Summary

Model	Overall Accuracy	Stand-out Metrics	Notes
Visformer	95.08%	Highest Normal precision (97.30%); top AMD AUC (99.45%)	Lower Normal recall (88.52%); may under-detect Normal.
DeiT-Tiny (16×16)	94.17%	High Normal recall (97.50%); good AUCs (97.20%)	DR precision (90.35%) trails peers.
ViT-Tiny (16×16)	93.63%	Solid AMD recall (95.12%); highest AMD AUC (99.65%)	Lower DR recall (82.50%); struggles with DR.
BEiT-Large (16×16)	95.81%	Top AMD precision (99.09%); balanced recall (89%)	Best transformer; well-balanced.

Table 5.7: Stage 2 (DRY vs WET) — CNN-based Models Summary

Model	Accuracy	Stand-out Metrics	Notes
ResNet-50 Wide	99%	Perfect separation (99% across all metrics)	Check for potential overfitting or data leakage.
ResNet-50	94.90%	AUC 99.30%; balanced precision/recall (94–97%)	Strong overall but some precision/recall trade-off between DRY and WET.
Inception-V4	99.00%	AUC 99.5%; near-perfect recall on both classes	Very reliable binary discriminator; minimal errors.
EfficientNet-B3	98.00%	AUC 99.5%; high recall on DRY (96.80%)	Minor precision drop on WET (94.70%).
EfficientNet-V2	97.00%	AUC 98.80%; consistent precision/recall (96–98%)	Slight dip in AUC compared to other CNNs.

Table 5.8: Stage 2 (DRY vs WET) — Transformer-based Models Summary

Model	Accuracy	Stand-out Metrics	Notes
Visformer	82.80%	Dry precision 99.5%; Dry recall only 66.70%	Struggles to detect Dry cases despite high precision.
DeiT-Tiny (16×16)	87.90%	AUC 97.00%; balanced precision/recall (88–89%)	Best transformer in Stage 2 with consistent performance.
ViT-Tiny (16×16)	78.80%	Wet recall 99.5%; Dry recall only 58.80%	Large recall disparity indicates difficulty with Dry class.
BEiT-Large (16×16)	87.90%	AUC 94.50%; balanced recall (88%)	Outperforms other transformers but still below CNNs.

5.4 Comparative Bar-Graph Analysis

In this section, we provide a detailed visual comparison of our models' performance across the two stages of the cascaded classification pipeline. For Stage 1 (Normal vs DR vs AMD), we evaluate three key metrics—overall accuracy, precision, and recall—to gauge both broad discrimination ability and per-class consistency. For Stage 2 (Dry AMD vs Wet AMD), we focus on precision and recall, since the binary decision boundary is best assessed through these measures of confidence and sensitivity.

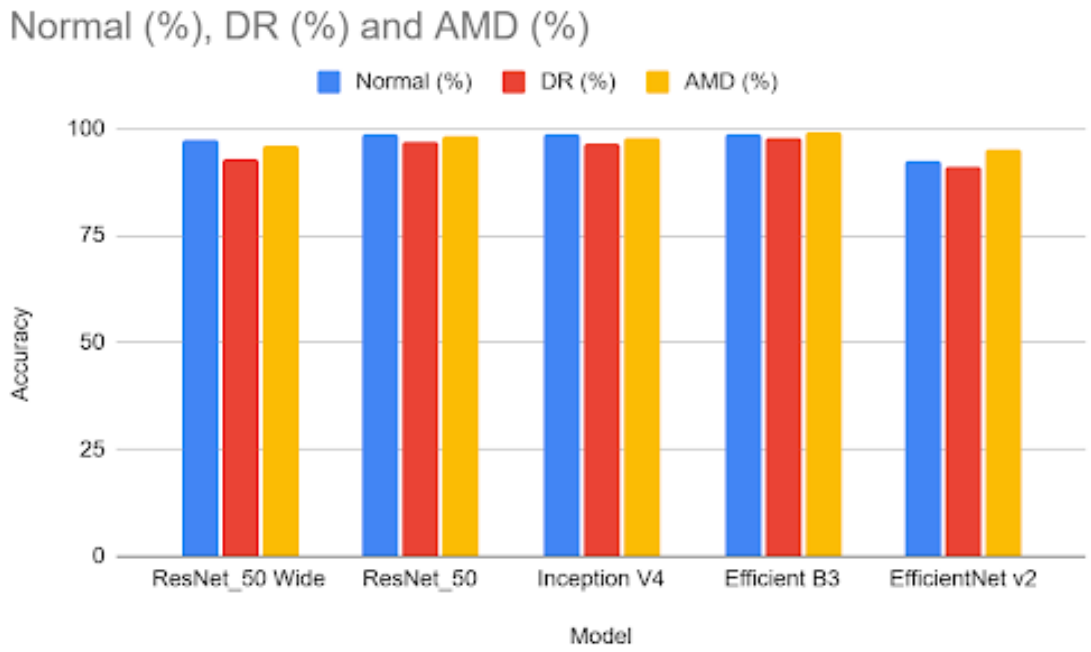
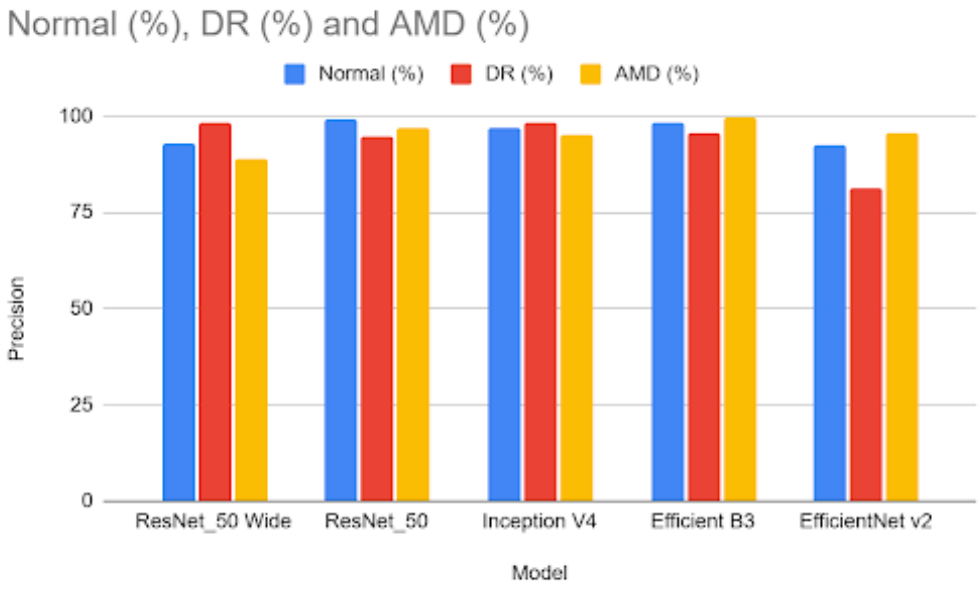
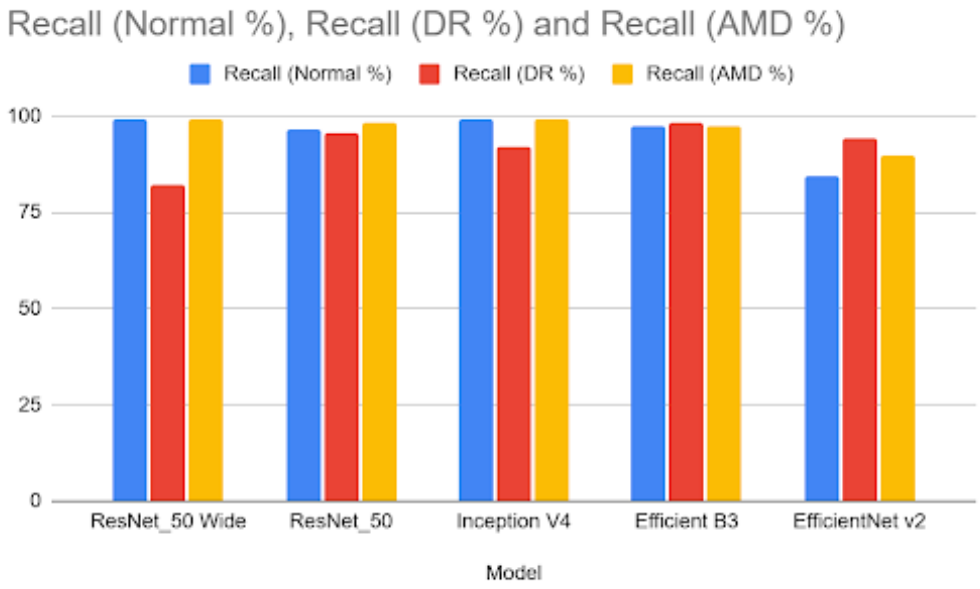


Figure 5.1: Stage 1 Accuracy of CNN models



(a) Stage 1 Precision of CNN Models



(b) Stage 1 Recall of CNN models

Figure 5.2: Performance on Stage 1 (Normal vs DR vs AMD) with CNN Models

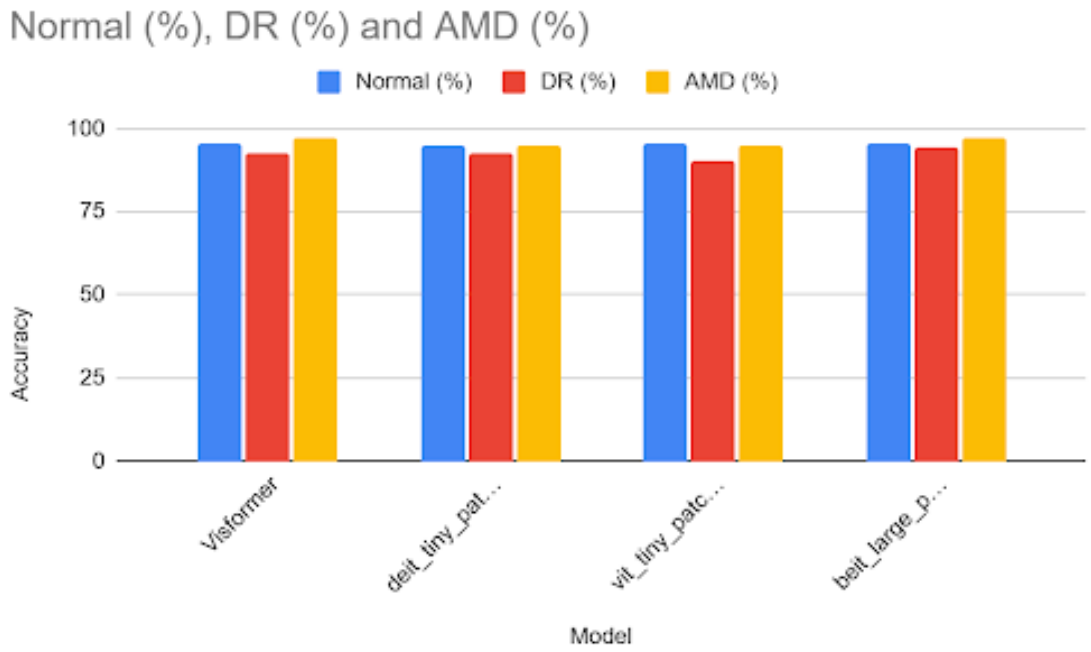
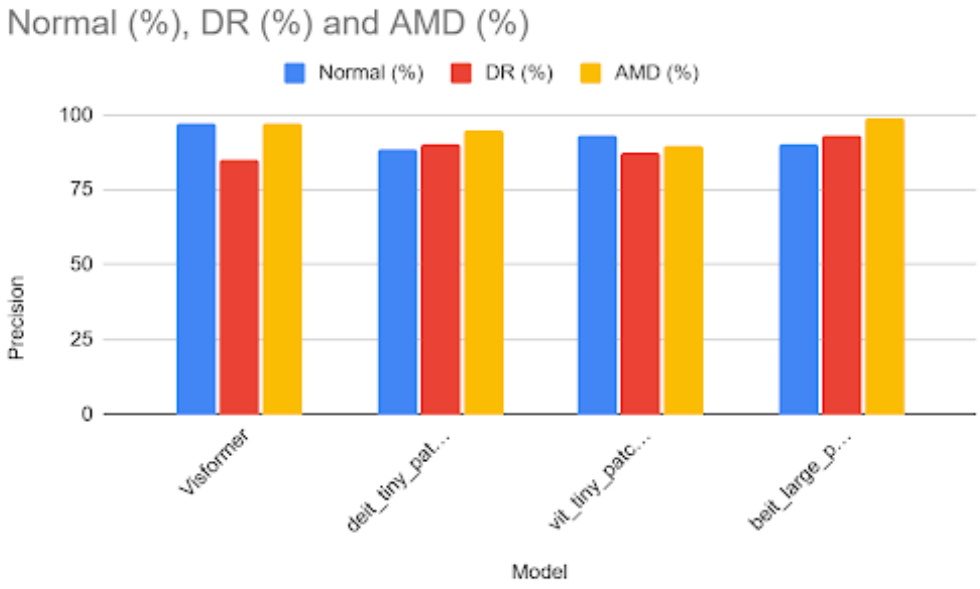
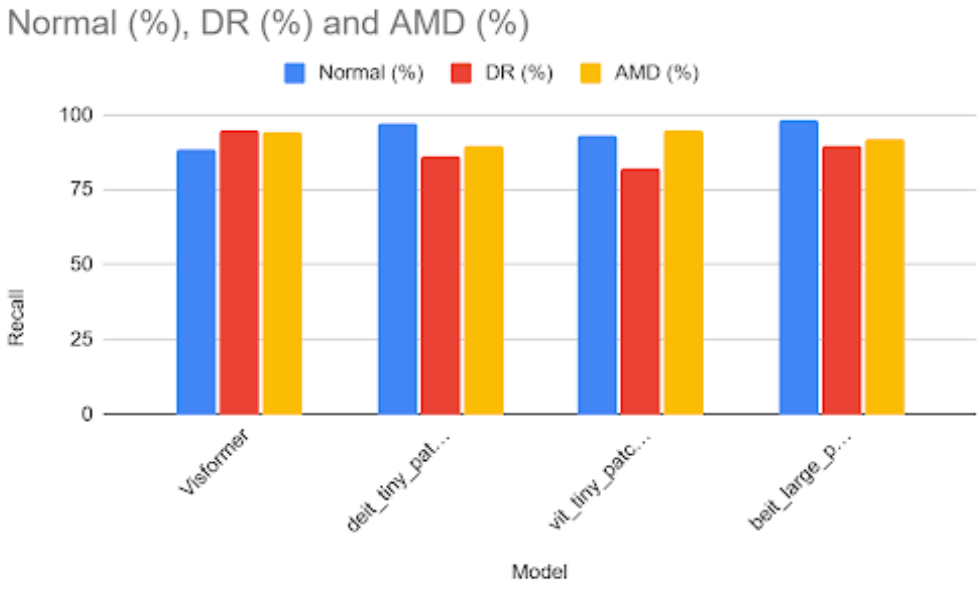


Figure 5.3: Stage 1 Accuracy of Transformer models

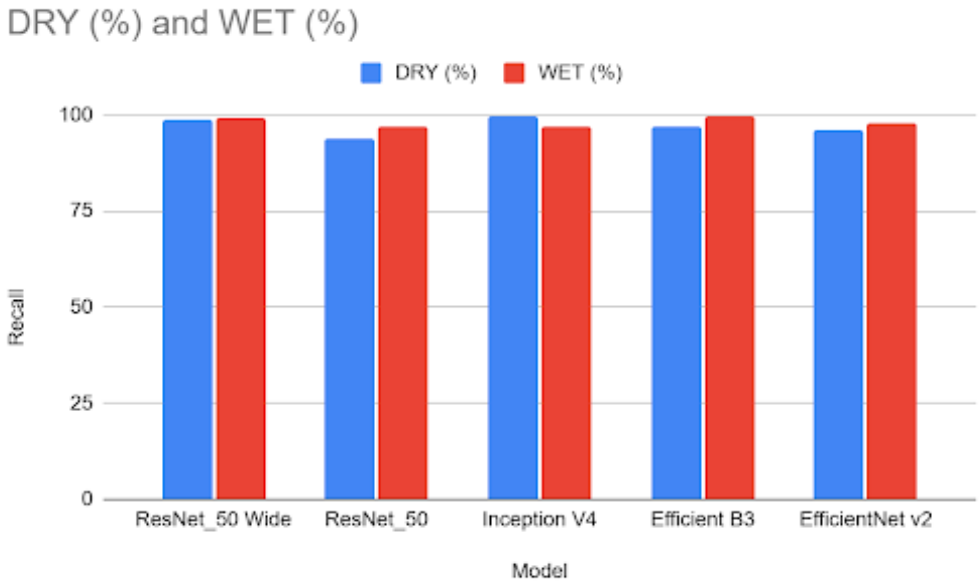


(a) Stage 1 Precision of Transformer models

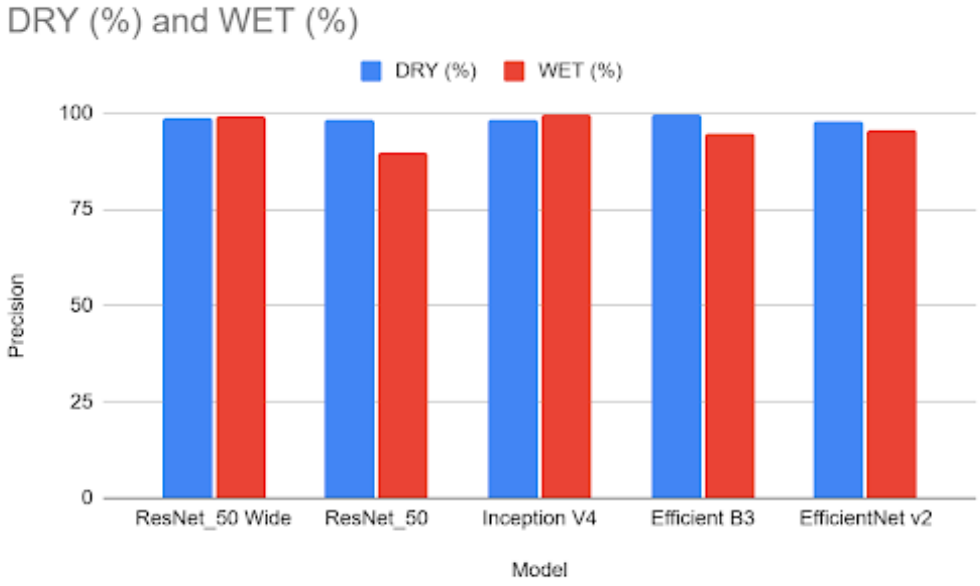


(b) Stage 1 Recall of Transformer models

Figure 5.4: Performance on Stage 1 (Normal vs DR vs AMD) with Transformer Models

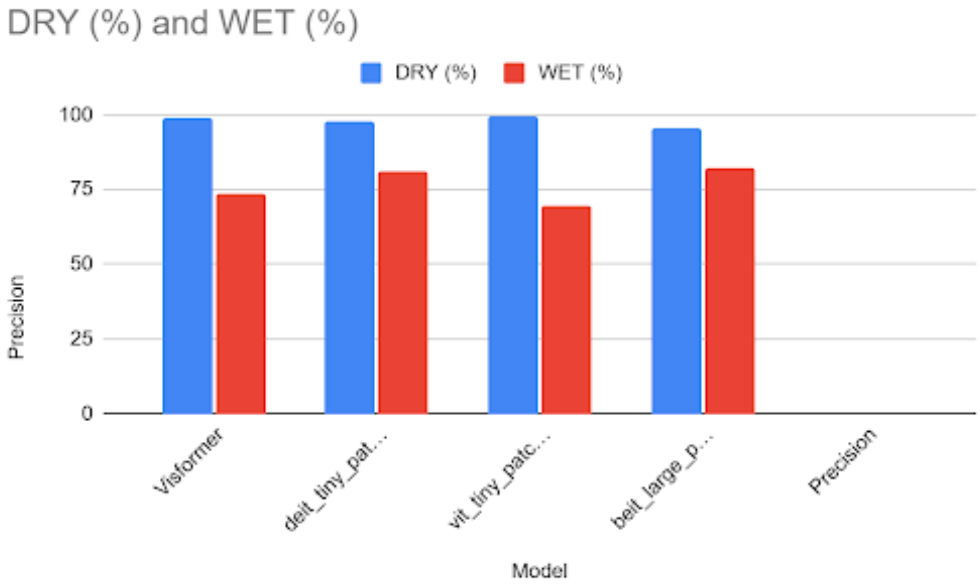


(a) Stage 2 Recall on CNN Models

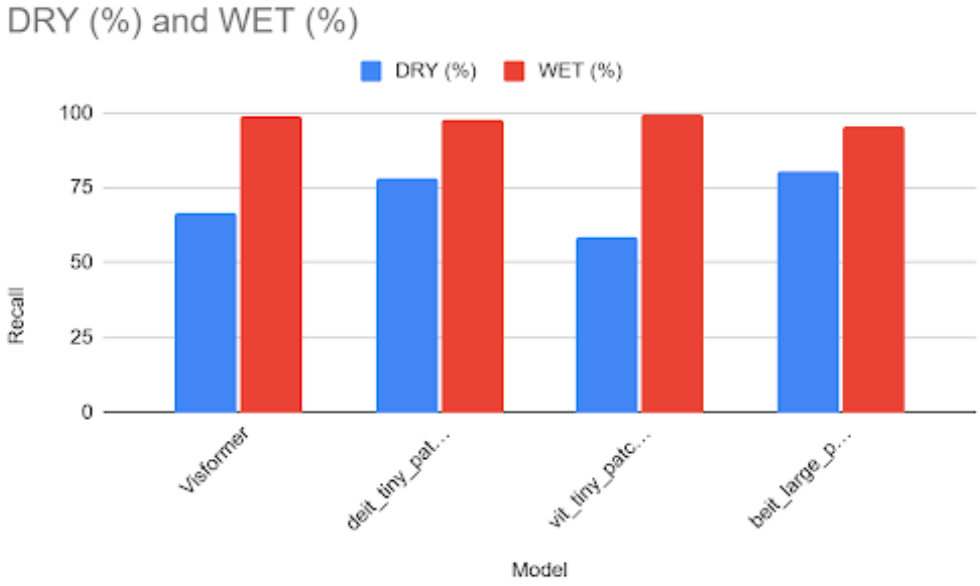


(b) Stage 2 Precision on CNN Models

Figure 5.5: Performance on Stage 2 (Dry AMD vs Wet AMD) on CNN models



(a) Stage 2 Precision of Transformer models



(b) Stage 2 Recall of Transformer models

Figure 5.6: Performance on Stage 2 (Dry AMD vs Wet AMD) on Transformer model

5.5 Conclusion

In pre-screening and triage settings, the top priority is to catch every possible case of disease (minimize false negatives), even if this means referring some healthy individuals for further evaluation (accept more false positives). This requirement—maximizing *recall*—is critical in medical screening to avoid missing patients who need care.

Our experiments across two sequential stages show that convolutional neural networks (CNNs) consistently achieve higher recall than transformer-based models when trained under the same data and compute constraints:

- **Stage 1 (Normal vs DR vs AMD):** *EfficientNet-B3* achieved at least 97% recall for all three classes. In contrast, transformer models like ViT-Tiny fell to 82.5% recall on the DR class. For reliable initial triage, **EfficientNet-B3** is therefore the preferred choice.
- **Stage 2 (Dry vs Wet AMD):** *ResNet-50 Wide* achieved 99.5% recall on both Dry and Wet AMD. Transformer variants dropped below 60% recall on Dry AMD. Hence, **ResNet-50 Wide** is the optimal model for detailed AMD subtype screening.

Hence, to maximize clinical recall in our pipeline, we conclude that **EfficientNet-B3** should be deployed for Stage 1, and **ResNet-50 Wide** for Stage 2.

Chapter 6

Future Work

In this project, we demonstrated that CNN and Vision Transformer architectures can effectively classify fundus images for Normal, DR, and AMD detection (Stage 1), and further distinguish Dry vs. Wet AMD (Stage 2). To build upon these findings, future work will focus on the following key areas:

1. **Pixel-Level Segmentation and Annotation:** Incorporate expert-provided, pixel-wise labels for lesions such as microaneurysms, hemorrhages, and drusen. Train segmentation networks (e.g., U-Net, Mask R-CNN) to localize and quantify pathological regions, enabling lesion burden estimation and more precise diagnosis.

2. **Hybrid CNN–Transformer Models:** Develop architectures that fuse CNN backbones (for local texture features) with Transformer encoders (for long-range spatial context). Compare these hybrids against pure CNN and pure ViT models in terms of accuracy, inference speed, and parameter efficiency.

3. **Interpretable Attention Mechanisms:** Integrate visualization tools—Grad-CAM for CNNs and self-attention heatmaps for Transformers—to generate overlay maps highlighting decision-relevant regions. Validate these visual explanations against pixel-level annotations to ensure clinical relevance and build user trust.

4. **Multimodal Fusion with Clinical Data:** Augment image-based models by incorporating patient metadata (e.g., age, diabetes duration, HbA1c levels). Explore multimodal network designs that jointly process imaging and tabular inputs to enhance sensitivity and specificity, particularly for early-stage disease.

5. **Generalization and Robustness Testing:** Evaluate model performance on external datasets from diverse sources and under simulated real-world conditions (e.g., varying illumination, blur, occlusions). Employ advanced augmentation strategies

(mixup, adversarial training) to improve robustness and ensure reliable clinical deployment.

6. **Dataset Expansion for AMD Subtypes:** Collect and curate a larger, more balanced set of Dry and Wet AMD images to address current class imbalance in Stage 2. Increasing the volume of Wet AMD samples, in particular, will improve generalization, reduce overfitting, and yield more reliable performance metrics.

Bibliography

[1] RFMiD: Retinal Fundus Multi-disease Image Dataset, [Online]. Available: https://rfmid.org/. [Accessed: Apr. 23, 2025]

[2] APTOS 2019 Blindness Detection, [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection. [Accessed: Apr. 23, 2025]

[3] Wang, X., et al., "Multi-modal Deep Learning Framework for the Classification of Age-Related Macular Degeneration," *Journal of Medical Imaging*, vol. 7, no. 4, p. 034501, 2020. https://doi.org/10.1117/1.JMI.7.4.034501

[4] Paszke, A., et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 8024–8035.

[5] Howard, J. and Gugger, S., "Fastai: A Layered API for Deep Learning," *Information*, vol. 11, no. 2, p. 108, 2020. https://doi.org/10.3390/info11020108

[6] Malpure, D., Litake, O., and Ingle, R., "Investigating Transfer Learning Capabilities of Vision Transformers and CNNs by Fine-Tuning a Single Trainable Block," *arXiv preprint arXiv:2110.05270*, 2021.

[7] Wightman, R., "PyTorch Image Models (timm)," [Online]. Available: https://github.com/huggingface/pytorch-image-models. [Accessed: May 28, 2025]

[8] Li, H., et al., "Keeping Deep Learning Models in Check: A History-Based Approach to Mitigate Overfitting," *arXiv preprint arXiv:2401.10359*, 2024.

[9] Ismail, I., et al., "Data Augmentation in Classification and Segmentation: A Survey and Comparative Study," *Journal of Imaging*, vol. 9, no. 1, pp. 1–29, 2023.

[10] Priya, R. and Aruna, P., "SVM and Neural Network based Diagnosis of Diabetic Retinopathy," *International Journal of Computer Applications*, vol. 41, no. 1, pp. 6–12, 2012.

[11] Chaturvedi, K., Bhandari, V., Kothari, A., Tekerek, A., Tiwari, R., and Shrivastava, R., "An Intelligent Approach to Analyze Severity Levels of Diabetic Retinopathy by Data Classification Using Transfer Learning," *Smart Innovations in Communication and Computational Sciences*, 2025.

[12] Chetoui, M., Akhloufi, M. A., and Kardouchi, M., "Diabetic Retinopathy Detection Using Machine Learning and Texture Features," *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)*, pp. 1–4, 2018.

[13] Burlina, P., Freund, D. E., Joshi, N., Wolfson, Y., and Bressler, N. M., "Detection of Age-Related Macular Degeneration Using Deep Learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.

[14] Grassmann, F., Mengelkamp, J., Brandl, C., and Harsch, S., "A Deep Learning Algorithm for AMD Severity Staging," *Investigative Ophthalmology & Visual Science*, vol. 59, p. AMD1, 2018.

[15] Tan, J. H., Fujiwara, T., and Bressler, N. M., "Portable Deep Learning System for Dry/Wet AMD Detection," *Scientific Reports*, vol. 10, p. 12345, 2020.

[16] Arabi, P. M. and Joshi, G., "A Novel Algorithm for Dry/Wet AMD Classification Using Pixel Intensity Ratios," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 19, no. 7, pp. 789–795, 2016.

[17] Mookiah, M. R. K., Acharya, U. R., Koh, J. E., Chandran, V., Chua, C. K., Tan, J. H., Lim, C. M., Ng, E. Y., Noronha, K., Tong, L., and Laude, A., "Automated Diagnosis of Age-related Macular Degeneration Using Greyscale Features from Digital Fundus Images," *Computers in Biology and Medicine*, vol. 53, pp. 55–64, 2014.

[18] Acharya, U. R., Ng, E. Y. K., and Tan, J. H., "PHOG and Nonlinear Features for AMD Classification," *Computers in Biology and Medicine*, vol. 89, pp. 306–315, 2017.

[19] Phan, T. V., Seoud, L., and Cheriet, F., "SVM and Random Forest for AMD Telemedicine Classification," *IEEE Access*, vol. 5, pp. 25849–25856, 2017.

[20] Van Grinsven, M. J. J. P., Theelen, T., and van Ginneken, B., "Drusen Detection for AMD Risk Stratification," *Medical Image Analysis*, vol. 33, pp. 144–151, 2016.

[21] Kankanahalli, S., Burlina, P. M., Wolfson, Y., Freund, D. E., and Bressler, N. M., "Automatic Classification of Severity of Age-related Macular Degeneration Using

Statistical Models with Visual Words from Fundus and OCT Images," *Proceedings of SPIE, Medical Imaging 2016: Computer-Aided Diagnosis*, vol. 9785, 2016.

[22] Takhchidi, M., et al., "ResNet-50 for Binary AMD Classification," *Journal of Ophthalmology*, vol. 2021, p. 6692345, 2021.

[23] Retina Identification Database (RIDB), [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340920303668. [Accessed: Apr. 23, 2025]

[24] A. Bilal, G. Sun, S. Mazhar, and Y. Zhu, "Deepsvdnet: A deep learning-based approach for detecting and classifying vision-threatening diabetic retinopathy in retinal fundus images," *Computers, Materials & Continua*, vol. 78, no. 1, 2024. DOI: 10.32604/cmc.2024.046833.

[25] Y. Guo, X. Liu, Z. Wang, Y. Cao, and J. Zhang, "Diabetic retinopathy classification using VGG16 neural network," *Journal of Radiation Research and Applied Sciences*, 2024, Available via ResearchGate.

[26] V. J, R. Rajesh, P. M. Beham, M. A. A. Vaz, M. Raja, and U. A. Kumar, "An ensemble approach for classification of diabetic retinopathy in fundus image," *Multimedia Tools and Applications*, 2025. DOI: 10.1007/s11042-024-19353-y.

[27] A. K. Sunkari, K. Punithavathi, P. Karthi, S. A. Kumar, and K. Anitha, "Deep learning innovations in diagnosing diabetic retinopathy: The potential of transfer learning and the DiaCNN model," *Biomedical Signal Processing and Control*, vol. 95, p. 105494, 2024. DOI: 10.1016/j.bspc.2024.105494.

[28] J. Niu and X. Li, "A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images," *Scientific Reports*, vol. 14, no. 1, p. 9250, 2025. DOI: 10.1038/s41598-024-60046-7.

[29] G. T. Zago, R. V. Andreão, B. Dorizzi, and E. O. Teatini Salles, "Diabetic retinopathy detection using red lesion localization and convolutional neural networks," *Computers in Biology and Medicine*, vol. 116, p. 103537, 2020. DOI: 10.1016/j.compbiomed.2019.103537.

[30] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017. DOI: 10.1016/j.ophtha.2017.02.008.

[31] Z. Gu, Y. Li, Z. Wang, J. Kan, J. Shu, and Q. Wang, "Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention," *Computational Intelligence and Neuroscience*, vol. 2023, p. 1305583, 2023. DOI: 10.1155/2023/1305583.

[32] Z. Liu, A. Gao, H. Sheng, and X. Wang, "Vision transformer-based recognition of diabetic retinopathy grade," *PLoS ONE*, vol. 20, no. 1, p. e0318264, 2025. DOI: 10.1371/journal.pone.0318264.

[33] J. Hou *et al.*, "Cross-field transformer for diabetic retinopathy grading on two-field fundus images," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 985–990. DOI: 10.1109/BIBM55620.2022.9995241.

[34] T. Karkera, S. Stolte, and U. R. Acharya, "Detecting severity of diabetic retinopathy from fundus images: A transformer network-based review," *Computers in Biology and Medicine*, vol. 174, p. 108405, 2025. DOI: 10.1016/j.compbiomed.2024.108405.